IDENTIFYING OPPORTUNITIES TO IMPROVE CONTENT MODERATION

A Dissertation
Presented to
The Academic Faculty

By

Shagun Jhaver

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology

May 2020

IDENTIFYING OPPORTUNITIES TO IMPROVE CONTENT MODERATION

Approved by:

Dr. Amy Bruckman, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Eric Gilbert, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Neha Kumar
Sam Nunn School of International Affairs and the School of Interactive Computing
Georgia Institute of Technology

Dr. W. Keith Edwards
School of Interactive Computing
Georgia Institute of Technology

Dr. Scott Counts
Social Technologies Group
Microsoft Research

Date Approved: March 4, 2020

In loving memory of my mom.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the generosity and support of many, many people. First, I would like to thank my advisors, Amy Bruckman and Eric Gilbert. They provided me intellectual and emotional support at every step of the way. Our weekly meetings always left me with a clearer focus and joy for my research. I will always cherish these meetings. Most importantly, Amy and Eric showed me how to be a good advisor through their example. Since early on in my PhD, I made a conscious effort to absorb the leadership lessons and values that this collaboration has offered. I look forward to practicing these lessons in my future roles as a teacher, advisor, mentor and leader.

I am also deeply indebted to an extraordinary group of mentors: Neha Kumar, Michaelanne Dye, Stevie Chancellor, Scott Counts and Munmun De Choudhury. Their thoughtful advise has been helpful in shaping my research. I also consider myself fortunate to have made friends who I will cherish forever: Koustuv Saha, Benjamin Sugar, Eshwar Chandrasekharan and Bahador Saket. Additionally, I could not have asked for a more supportive group of lab-mates for this journey: Sucheta Ghoshal, Julia Deeb-Swihart, Darren Scott Appling and Jane Im. To each of you: Thank you!

I have also been fortunate to have an amazing network of friends outside my department. To my housemates: Kashyap Mohan, Sloane Tribble, Akshay Sakariya, Swapnil Srivastava and Amber Chowdhary, your constant support has been crucial in maintaining my spirits during the tough times. I have been much enriched by our informal brainstorming sessions. I have also savored every minute of our board games together!

Above all, I am incredibly grateful to my extended family in India and the US. To my grandmother, sister, uncles, aunts, cousins, and friends, you have provided me the emotional and financial support I so desperately needed to continue my journey along this chosen path. I am deeply thankful for the support that I continue to receive from all of you.

Finally, I would like to thank hundreds of individuals who participated in my studies.

Without their voluntary support, this work would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis contributes a nuanced understanding of the challenges inherent in the design and implementation of fair and efficient content moderation systems. Using large-scale data analyses, participant observations, survey data and in-depth qualitative interviews, this research describes the perspectives and practices of different stakeholders - users who suffer online harassment, individuals whose posts get removed, people who rely on blocking tools to censor others, and community managers who volunteer time to regulate content. This work provides theoretical and practical guidelines for moderating against online harassment without impinging on free speech, for designing solutions that incorporate the needs of different user groups, and for adopting automated moderation tools that provide explanations of their decisions and that remain sensitive to localized contexts.

CHAPTER 1
INTRODUCTION

Today, millions of individuals use social media platforms like Facebook, Twitter and Reddit on a daily basis. Although these platforms don't create their own original content, they allow users to host their content and share information and ideas with one another. A fundamental service these platforms provide is that they determine for each user posting, whether that posting will be allowed to stay online or be removed, how prominently it will be shown if it is allowed, and what additional measures are taken if it is removed. These decisions are made by a sociotechnical system called content moderation, which comprises site administrators, commercial and/or voluntary moderators, end-users and automated systems. Understanding how this system works and how users respond to it is at the heart of my dissertation.

Studying content moderation systems is important because decisions about allowing or removing different types of user content, the reasoning behind these decisions, and the ways in which they are implemented have important consequences. For one, these moderation processes ensure that the limited attention of users is not exhausted by unworthy posts. More importantly, these processes influence which groups' voices get heard, and whether minorities and other vulnerable groups feel comfortable participating online. Moderation shapes whether online communities allow users with opposing views to constructively interact with one another or whether these communities become echo chambers in which users merely reinforce one another's beliefs. As more and more of our conversations, cultural production and public discourse move online and as social media platforms continue to advance their cultural, economic and political power, it is incumbent upon us to examine how their moderation systems work. Studying how the current moderation apparatus are built and attending closely to why they fail to address problems like online harassment or misinformation campaigns can provide us insights into the deficiencies of current content moderation policies and the sociotechnical mechanisms that enforce those policies. It can also inform the design of new solutions that may address these deficiencies and encourage prosocial outcomes.

Moderation is an essential part of what platforms do. Yet, it is largely hidden from the public in order to maintain the illusion of an open platform and to avoid any legal or cultural responsibility (Gillespie, 2018a). The ways in which social media platforms moderate today have begun to settle in as the established and accepted ways to handle what we post online. But what type of moderation systems do we really want? What does fairness in content moderation mean to us as the end-users? These questions are important to consider before we can think about how platforms can affect the dynamics of society in positive ways. Moderation systems are also notoriously opaque, revealing little about how they are constituted or how they operate, and having minimal external accountability. Therefore, highlighting the internal processes and perspectives of different stakeholders in these systems is crucial to understanding their limitations and to building better solutions.

Although it is easy to blame platforms for the internet's troubles — fake news and misinformation campaigns, trolling and harassment, virulent misogyny and online radicalization, we should recognize that content moderation is *hard*. Moderators on these platforms need to continually monitor the incoming stream of posts. They often have to make difficult decisions about how to balance offense and importance, how to reconcile competing value systems, how to distinguish political discourse and violent rhetoric, and how to account for individual differences in gender, race and sexuality (Gillespie, 2018a). Should explicit but socially valuable material such as images from wars be allowed? Should mothers posting breastfeeding photos be blocked? How much sex is too much sex? These questions bring up age-old debates about the proper boundaries of public discourse as well as raise new ones. Platforms have to make significant investments into the human, technical and financial resources to execute these decisions in ways that avoid public outrage and legal interventions. Yet, platform initiatives, intended to save costs, can often end up making unjust or inconsistent moderation decisions. It is, therefore, imperative to explore how new policy and technological interventions change how moderation systems work and how they affect community dynamics. Such efforts can inform other existing (as well as future) platforms how to make difficult moderation decisions as well as what to expect from potential design and policy interventions.

While the public recognition of the importance of content moderation has grown in recent years, content moderation is not a novel process. Since the early days of the internet, scholars as well as community managers have deliberated over how to manage content online and how to enable productive conversations among end users (Dibbell, 1994; Bruckman et al., 1994). Although modern social media platforms operate in new contexts, they still struggle with some of the same problems that afflicted the early incarnations of online communities. This raises the question — Despite decades of research on online communities, why have we still *not* solved the problems of content moderation? Why is it still hard to deal with "trolls" (Herring et al., 2002) or make subjective decisions that distinguish controversial speech from online harassment (Chapter 3)? It turns out that content moderation is as much a social problem as it is a technical problem. Regulating online spaces in ways that address the evolving social challenges and balance the needs of different user groups can be considerably difficult. Just as importantly, as social media platforms grow increasingly large, moderation strategies that work for smaller online communities may prove to be too costly, thereby requiring novel moderation solutions that are efficient as well as acceptable to end-users. Given these challenges, human-centered research on content moderation is important to address the deficiencies in the current systems.

1.1 Research Framing

My dissertation consists of studies around five important aspects of content moderation. First, in Chapter 3, I discuss the problem of *online harassment* and the challenges of moderating against online harassment. Second, in Chapter 4, I present the use of *third-party moderation systems* and consider what we can learn by analyzing them. Third, in Chapter 5, I explore how content moderation is enacted in practice, focusing on the use of *automated tools* for content moderation, and identifying the challenges of using those tools. Fourth, in Chapter 6, I analyze what *fairness* in content moderation means from the perspectives of users whose posts are removed. Finally, in Chapter 7, I discuss *transparency* in content moderation, focusing on evaluating the effects of transparent moderation on long-term user behaviors.

In the rest of this section, I will briefly introduce each of these five aspects of content moderation, highlighting my reasons for exploring them in-depth.

1.1.1 Online Harassment

One of the primary goals of content moderation is to detect and remove instances of online harassment. It is vital that moderation systems efficiently address online harassment because it has emerged as a significant social problem over the past few years. A recent Pew research study found that 41% of American adults have suffered online harassment (Duggan, 2017). Minorities and other vulnerable groups such as LGBT users are frequently targets of online abuse. Therefore, efficient anti-abuse moderation strategies are critical to implement so as to ensure that minority groups feel comfortable participating online and they have a voice in the public sphere.

Although preventing online harassment seems to be a clear, tenable goal at first glance, it can be surprisingly difficult to implement in practice. This is because the question of what should be considered online harassment is often subjective. Overtly harassing posts such as rape threats and death threats are considered offensive and unacceptable by most users and can be removed with little hesitation, but how should one adjudicate cases where the offense is much milder? How about instances where an offensive post is made as a political act or in response to another offensive post? How should we moderate posts that may be offensive to one user group but is acceptable to others? Reflection on such questions begins to highlight the challenges of implementing content moderation against online harassment in ways that are acceptable to all users.

While the need to address online harassment is gaining wider recognition, many Americans also fervently believe in the ideal of ‘freedom of speech’. Section 230 of the U.S. telecommunications law provides immunity to internet intermediaries to regulate user-generated content however they see fit and does not consider them responsible for upholding their users’ freedom of speech (Gillespie, 2018a). Still, many users feel that they should be allowed to post whatever they want on these sites. Since there are some users who desire absolute freedom of expression and others who expect the sites to moderate online harassment, community managers are in a dilemma — if they implement content moderation in such a fashion that most posts are allowed to stay online, users who encounter offensive comments may get frustrated and leave the community. On the other hand, if the community managers implement strict moderation, users whose comments are removed may feel they are being censored and they may leave the community. Therefore, managers have to strike a fine balance in their work so that their decisions are acceptable to most users and do not result in mass user withdrawals from the site.

To better understand the challenges involved in finding this balance and to analyze the boundaries between controversial speech and online harassment, I studied a specific online community called Kotaku in Action. This community had been accused of perpetrating online harassment in the news media, but its members deny such accusations while proclaiming themselves as free speech proponents. Interviewing members of this community, I gained theoretical insights on moderating against online harassment that I present in Chapter 3.

1.1.2 Third-party Moderation Systems

Inefficient content moderation can be frustrating for users. In an effort to address these frustrations, social media platforms like Twitter, Facebook and Reddit have designed a variety of tools for users to resist being harassed and to control what type of posts they see on the site. For example, Twitter users can *block* an offensive account so that they don't view any posts from that account. But do these tools work for everyone? What gaps exist between the needs of users and the affordances provided by social media platforms when addressing the problem of online harassment? Answering these questions is important to understanding the challenges of content moderation.

As I will discuss in detail in Chapter 4, many users do not find the default moderation tools provided by social media websites sufficient to meet their needs. This has led to the creation of third-party moderation mechanisms that provide users greater control to manage the content they receive. One such mechanism is Twitter blocklists, a tool that allows Twitter users to pre-emptively block with a few clicks all accounts on a community-curated or algorithmically generated list of block-worthy accounts. Studying the use of this tool from the perspectives of (1) users who subscribe to blocklists and (2) users who are blocked on these lists, I found that although the use of this tool helps reduce the problem of online harassment for many users, it also leads to unintentional blocking of many users who feel disconnected from important groups. I discuss the implications of these outcomes and present other details of this study in Chapter 4.

1.1.3 Automated Tools for Moderation

It is important to understand the perspectives of users who engage with content moderation systems so that we can identify the deficiencies of the current systems and begin to improve them. At the same time, it is also crucial to understand how individuals who are in charge of managing moderation systems do their job. These individuals are called *moderators*, and they can be either commercial workers hired by social network sites (e.g., for Twitter and Facebook) or they can be voluntary users of their communities (e.g., for Reddit and Facebook Groups).

One of the primary challenges many moderators face is the sheer volume of content they need to moderate. For example, many popular Reddit communities have millions of subscribers and thousands of new postings every day. How do moderators make nuanced moderation decisions on so much content? It turns out that moderators often deploy automated tools to offload part of their work. In an effort to better understand the use of these tools, I focused on Reddit Automoderator, a widely used automated moderation tool on Reddit. My findings show that Automoderator makes many easy decisions automatically so that human moderators are only responsible for adjudicating cases that are harder and require more subjective interpretation. Yet, these human-machine collaborations are far from perfect. For example, Automoderator often make mistakes that human moderators need to correct. These tools also require regular configuration and updating that many moderators find technically challenging.

The use of AutoModerator is an illustrative example of human-technology partnerships wherein evolving technologies are actively shaping the lives of workers who, in turn, are shaping those technologies. Building upon the findings of this study, I describe the challenges that moderation systems can expect to face as they adopt novel automated tools that operate in harmony with human workers. These challenges include a consideration of the skills required to use new automated tools. I present the details of this study in Chapter 5.

1.1.4 Fairness in Content Moderation

End-users are the central actors in online social systems. Sites like Twitter and Reddit don't usually create their own content. Instead, they rely on a constant stream of user-generated content (Gillespie, 2018a). Therefore, end-users are not just consumers who bring in the ad revenue to sustain these platforms but they are also the content creators. Given the platforms' reliance on end-users for content creation and ad revenues, it is crucial for these platforms to have users who are invested in the online community and who feel valued for their content contributions.

Although many users on these platforms create information goods that are appreciated by the community, there are others whose posts are promptly removed by the community managers before they can be seen by the community. We do not know what happens to users after they invest time in creating content only to have it discarded. It is also unclear how the different elements of submission process (e.g., the existence of community guidelines) and the subsequent removal process (e.g., whether or not the user was provided a removal reason) affect users.

To investigate this space, I conducted a study of Reddit users who experienced content removals, asking them their views on the fairness of content removals and their attitudes about posting in the future. Through understanding the concerns and experiences of moderated users, my work opens up opportunities for identifying and nurturing users who have the potential to become valuable contributors in the community. I discuss this study in detail in Chapter 6.

1.1.5 Transparency in Content Moderation

How moderation happens on a community can have an enormous influence on users. Moderation can affect whether users choose to continue being a member of that community, how they engage with other users, and how close-knit the community becomes. One crucial aspect of moderation is the level of transparency about moderation actions. Strategic or managed transparency can be a beneficial means to engender trust among the users. However, the implications of promoting transparency in moderation are not obvious. On one hand, when explanations for content removals are made available, users have more opportunities to understand the social norms of a community. This may encourage them to become more productive members of that community in the future. On the other hand, if the users feel that the removal of their posting was not justified, it may discourage them from posting again on the community or it may provoke rebellious behavior such as offensive posts. Increasing transparency in moderation also requires more work on the part of moderators. For example, moderators have to invest more time and efforts to provide a correct reason for removing a post. Therefore, it is important to identify whether implementing transparency through providing the reasoning behind content removals is worth the cost of additional efforts on the part of moderators.

I investigated this question through a large-scale data analysis of Reddit posts, focusing on the ways in which providing removal explanations affects the future posting behavior of Reddit users. I use the empirical insights of this study to provide meaningful guidelines to community managers for designing transparency in their moderation processes. I also argue that content moderation should serve an educational rather than a punitive role. I present the results of this study in Chapter 7.

To sum up, my dissertation analyzes five important aspects of content moderation: (1) Distinctions between online harassment and controversial speech, (2) Use of third-party moderation tools, (3) Human-machine collaboration for content moderation, (4) Fairness of content moderation, and (5) Transparency in content moderation. For understanding content moderation, this *entire* picture is important – it is vital that we understand not just the perspectives of different user groups but also the conditions under which moderators work as well as the limitations of current moderation mechanisms. My work explores the current state of default and third-party content moderation systems, how they are used and implemented, and where they fall short. This holistic view of content moderation allows me to identify opportunities for improvements that are both meaningful for the users as well as feasible for the community moderators to implement.

1.2 Research Questions

My research aims to understand the processes that underlie content moderation systems and explore how different user groups experience and circumvent regulatory sanctions.

My goal is to establish the challenges of implementing content moderation and propose solutions that begin to address these challenges. Through five individual studies, I identify opportunities for improving content moderation, both from the perspectives of users as well as community managers.

High-level research questions include:

RQ 1: How can moderation systems improve the way they adjudicate instances of controversial speech? How do these systems distinguish passionate disagreements from harassment, and where do they draw the line between freedom of expression and censorship?

RQ 2: What are the limitations of currently offered default moderation systems for users who suffer online harassment? How does the use of third-party moderation mechanisms affect different user groups? What can we learn from the use of these mechanisms to improve content moderation?

RQ 3: How does the socio-technical system of content moderation work? How does the use of automated regulation mechanisms change the work of moderators? How can we enhance human-machine collaborations to attain better content moderation?

RQ 4: How do moderated users perceive content removals? In what ways do the contextual factors of post submissions and moderation processes, such as community guidelines and removal explanations, affect users' perceptions of fairness of content removals and their attitudes about posting in the future?

RQ 5: What type of removal explanations are typically provided to users? How does providing explanations affect the future posting activity and future post removals?

In order to answer these questions, I have conducted mixed-methods research, using a wide range of methods in my work. Findings from these five studies provide insights into the nuances of implementing efficient and acceptable content moderation and the steps that can be taken to improve the underlying processes. Table 1.1 summarizes these studies.

Table 1.1: Summary of Completed and Proposed Studies

Study	Research Questions	Data Sources
Conceptualizing Online Harassment: The Case of Kotaku in Action	- How can moderation systems improve the way they adjudicate instances of controversial speech? - How do these systems distinguish passionate disagreements from harassment, and where do they draw the line between freedom of expression and censorship?	- Participant observation on KiA subreddit. - Semi-structured interviews with KiA users.
Understanding the Use of Third-Party Moderation tools: The Case of Twitter Blocklists	- What are the limitations of currently offered default moderation systems for users who suffer online harassment? - How does the use of third-party moderation mechanisms affect different user groups? - What can we learn from the use of these mechanisms to improve content moderation?	- Network analysis of Twitter users. - Semi-structured interviews with (1) users who subscribed to blocklists; (2) Users who are blocked by blocklists.
Human-Machine Collaboration for Content Moderation: The Case of Reddit Automoderator	- How does the socio-technical system of content moderation work? - How does the use of automated regulation mechanisms change the work of moderators? - How can we enhance human-machine collaborations to attain better content moderation?	Semi-structured interviews with Reddit moderators of five large communities.
Understanding User Reactions to Content Removals: A Survey of Moderated Reddit Users	- How do moderated users perceive content removals? - In what ways do community guidelines and removal explanations affect users' attitudes about content removals and future postings?	- Survey of users whose content is removed.
Evaluating the Importance of Transparency in Moderation	- What type of removal explanations are typically provided to users? - How does providing explanations affect the future posting and future post removals?	- Large-scale content analysis.

1.3 Research Contributions

For this dissertation, I have focused on my studies on two popular platforms: Twitter, a microblogging provider, and Reddit, a multi-community discussion site. However, many findings and design suggestions from this dissertation should also transfer well to other similar platforms like Facebook (especially for Facebook Groups), Discord, Twitch and Instagram.

This dissertation offers the following research contributions:

- Using a case-study of the controversial site Kotaku in Action subreddit, I contribute a theoretical model for perceptions of controversial speech, dividing it into four categories – criticism, insult, public shaming and harassment. By listening to the perspectives of a group that is often overlooked – the users accused of online harassment, I highlight the sociocultural challenges of moderating content that is perceived differently by different user groups. My qualitative analysis of this online community offers valuable insights into how content moderation systems should approach the problem of moderating against online harassment without impinging on free speech. My findings reveal that moderation decisions should consider the intensity of the language used as well as the frequency of messages directed at a single target.

- This dissertation highlights that the current content moderation mechanisms don't satisfy the moderation needs of many user groups. Through conducting a case-study of Twitter blocklists, a third-party blocking mechanism used on Twitter, and focusing on what prompted the creation and popularity of this tool, I find the deficiencies of centralized moderation tools on Twitter. I show which users' needs are being met by this decentralized moderation mechanism. My work presents the benefits and limitations of using blocklists and shows how its use affects different user groups. I also use blocklists as a vehicle to explore the problem of online harassment and contribute a typology of tactics viewed as manifestations of online harassment by Twitter users.

I build on these findings to suggest opportunities for design that can help address the problem of online harassment while ensuring that users are not blocked unnecessarily or unfairly. Further, I suggest how user-centered design can help create moderation solutions that incorporate the specific needs of oppressed groups. I also explore the idea that the flip side of harassment is understanding across differences – the problems of harassment and lack of understanding are intertwined.

- Through interviews with 16 Reddit moderators, I peek inside the black box of content moderation on Reddit and reveal how the sociotechnical system of moderation works. I focus on how the moderators address the challenges of scale in moderating high-traffic communities by incorporating automated solutions. My work provides important insights into the problems that community managers can expect to face as they adopt novel automated solutions to help regulate the postings of their users. I also identify areas for improvement in the mixed-initiative moderation system of Reddit and propose ideas that can help improve the efficiency of content regulation.

- Using a survey of 907 users whose posts have been removed, I present a rich overview of moderated users, their concerns, and their interactions with various elements of the moderation process. This work offer recommendations for designers to build tools and community managers to adopt strategies that can help improve users' perceptions of fairness in content moderation and encourage them to become productive members of the community.

- Using a statistical analysis of 20 million Reddit posts, I show that provision of removal explanations is associated with lower odds of future submissions and future removals. Empirical findings from this dissertation contribute concrete guidelines for communities to design for explanations of content removals in their moderation strategies. I also highlight how automated tools can be particularly helpful in offering such explanations.

CHAPTER 2
RELATED WORK

This chapter describes the background and related work relevant to my thesis. At the heart of this work is the problem of content moderation, so I begin with providing a comprehensive background on different aspects of this complex multi-layered process. Next, I discuss online harassment, a growing social problem that content moderation aims to address, and that I focused on for two of my studies. Finally, I describe the issue of freedom of speech on internet platforms, focusing on the concepts that highlight the importance of establishing equitable content moderation systems.

2.1 Content Moderation

Over the last fifteen years, social media companies like Twitter, Facebook, YouTube, and Reddit have established themselves as the leading platforms in global content sharing (Collins, 2018; Klonick, 2017). Having hundreds of millions of users across the world, they account for a large part of the time that people spend on digital media. These platforms have also emerged as the prominent networked gatekeepers in the current information space because of their significant influence over information flows (West, 2017). Although they don't produce any new content themselves, they host and organize their users' content and make important choices about which material is accepted and which is refused, how the accepted material is shared and with whom, and in what ways can users interact (Gillespie, 2017a). Platforms enact information control at multiple levels and through various mechanisms, including site design, algorithmic curation (Eslami et al., 2015a; Gillespie, 2014; Jhaver, Karpfen, and Antin, 2018) and content moderation (West, 2017).

Perhaps the most explicit way in which platforms act as network gatekeepers is the sociotechnical process of content moderation (West, 2017). James Grimmelmann (2015) defines content moderation as “the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse”. Grimmelmann (2015) argues that content moderation has three important goals: (1) Creating productive communities that generate and distribute valuable information goods; (2) Increasing access to online communities and discouraging cutting people off from the knowledge the community produces; and (3) Lowering the cost of maintaining online communities by making as few demands as possible on the infrastructure as well as the participants. Different approaches to content moderation inevitably trade-off among these goals, especially as online communities increase in their size and traffic. For example, moderation may considerably prevent spam, harassment and other forms of abuse on a large community but that may drive up the costs of moderation to unacceptable levels. Indeed, moderating content in ways that are efficient, affordable and accepted by the community members can be significantly challenging for the community managers, as I will discuss later in this section.

Content moderation is not only a difficult but also an important task. It helps internet platforms present their best face to new users, advertisers, investors and the public at large (Gillespie, 2018a). Having taken on a curatorial role, these platforms “serve as setters of norms, interpreters of laws, arbiters of taste, adjudicators of disputes, and enforcers of whatever rules they choose to establish” (Gillespie, 2018a). The moderation decisions taken by these platforms determine what millions of users see and, just as importantly, what they don’t see. Therefore, content moderation plays a critical role in modern free speech and democratic culture. Indeed, the impacts of content moderation transcend online experiences and the question of freedom of speech. Prior research has suggested that failures in content moderation may cause irreversible professional damage or they may result in disabled or elderly users losing the ability to communicate with their family networks (West, 2018).

2.1.1 Challenges of Moderation Work

Alongside a growing recognition of the importance of content moderation, it is notable that in recent years, major social media platforms have repeatedly been criticized for how they moderate their content. Controversies over content policies and their implementations have raised questions over how platforms both fail to sufficiently remove disturbing material (e.g., fake news (Allcott and Gentzkow, 2017), alt-right trolls (Nagle, 2017; Romano, 2017), revenge porn (Citron and Franks, 2014; Vanian, 2017)) as well as censor material that is important to be circulated in the public sphere (e.g., ISIS beheadings (Kelion, 2013), war brutalities (Scott and Isaac, 2016), breastfeeding (Grenoble, 2013; West, 2017)).

While many internet platforms are supported by advertising as their business model and have an economic incentive to maximize site activity to increase ad revenue, they also have competing incentives to remove material that is likely to make their users uncomfortable, such as obscene or violent material, because keeping such content risks having users leave the platforms (Klonick, 2017). Consequently, social media platforms have to address the challenges of exactly when to intervene and in what ways. As platforms grow, they also face the challenges of scale in their role as content curators (Glaser, 2018; Madrigal, 2018). For example, many large Reddit communities have millions of subscribers, and their regulation systems have to process thousands of new comments posted every day. To take another example, Facebook's content regulation system must now process 510,000 comments posted every minute by 1.4 billion daily active users (Zephoria, 2018).

2.1.2 Composition of Moderation Systems

To address the challenges of processing high volumes of content efficiently, social media companies have created intricate and complex systems for regulating content at scale (Grimmelmann, 2015; Taylor, 2018). Such systems usually consist of a small group of full-time employees at the top who set the rules and oversee their enforcement, adjudicate hard cases, and influence the philosophical approach that the platforms take to govern themselves (Gillespie, 2017a). Platforms also employ a larger group of freelancers who work on contract with them and guard against infractions on the front line (Chen, 2014; Roberts, 2016). Finally, platforms rely on regular users to flag content that's offensive, harassing or violates the community rules (Crawford and Gillespie, 2016). Many platforms also depend on users to evaluate the worth of each comment through aggregating how users rate that comment. This is often referred to as distributed content moderation, and its benefits and limitations have been studied through prior research on moderation of the Slashdot website (Lampe, Johnston, and Resnick, 2007; Lampe and Resnick, 2004). This research shows that distributed moderation can enable civil participation on online forums (Lampe et al., 2014).

2.1.3 Role of Human Moderators

While the use of commercial content moderators is popular in the industry, not all platforms use paid staff to moderate their content. Many online platforms (e.g., Reddit, Wikipedia, Facebook Groups) largely rely on voluntary moderators who are given limited administrative power to remove unacceptable content and ban problematic users (Matias, 2016c; McGillicuddy, Bernard, and Cranefield, 2016). Reddit, in particular, contains more than a million independent communities, each with its own set of rules and volunteers who regulate content. Moderators are typically selected from among the users who are most actively involved in the community and are invested in its success (Matias, 2016c). Thus, they are well-suited to understand the local social norms and mores of the community (Diakopoulos and Naaman, 2011) and they engage in creating and enforcing local rules that establish the grounds for acceptable content. These voluntary moderators are not employees of the platforms but as communities get larger, moderators are often viewed by the users as representatives of the platforms (Matias, 2016a). Therefore, they constantly negotiate their positions with communities as well as with platforms (Matias, 2016c). Matias's work on Reddit moderation shows how community moderators on Reddit come to participate in collective action to protest and influence platform operators (Matias, 2016a).

Some prior research has explored the work of moderators. For example, Epstein and Leshed (2016) studied how moderators facilitate public deliberations on RegulationRoom, an online platform for policy discussions, and found a strong need to automate more of the moderators' work. Diakopoulos and Naaman (2011) studied regulation of online news comments on the website SacBee.com and found that although the moderators were concerned about handling a large volume of comments on the site, they were reluctant to outsource the control of content moderation. Moderators felt that outsiders might not have a locally meaningful understanding of the issues to be able to make appropriate moderation decisions in hard cases (Diakopoulos and Naaman, 2011). I will discuss in Chapter 5 how a similar inclination to maintain control affects the acceptance and use of Automod among Reddit moderators.

Emotional Labor of Moderation Work

A complementary line of research has focused on understanding the "emotional labor" (Hochschild, 1983) of moderation work (Kerr and Kelleher, 2015; Kiene, Monroy-Hernández, and Hill, 2016; Matias, 2016c; McGillicuddy, Bernard, and Cranefield, 2016; Roberts, 2014; Roberts, 2016). Sarah T. Roberts studied commercial content moderation (Roberts, 2014; Roberts, 2016) and found that many social media companies use the services of workers who are dispersed globally, pointing out that many of these workers suffer burnouts because of the routine, factory-like nature of the work of content moderation. Roberts also noted that the constant viewing of troubling content takes an emotional toll on the workers, and they resist discussing their work with friends and family to avoid burdening them (Roberts, 2014). In a similar vein, Kerr and Kelleher (2015) pointed out that such workers have to perform the emotional labor of enacting an "apolitical, culturally nuanced subjectivity" in their online work that may not align with their offline identity. Several recent media articles have also highlighted the emotional labor of moderation work (Buni, 2016;

Garcia, 2018; Renfro, 2016; Swearingen and Lynch, 2018). As I will elaborate in Chapter 5, I add to this literature by showing how automated content regulation on Reddit helps reduce the workload of moderators and allows them to avoid viewing large volumes of offensive content. I also explain the tradeoffs of this reduction in emotional labor via the use of automated tools, e.g., I discuss the automated removal of posts that may contain offensive language but are appropriate for the community in the context of their discussion.

2.1.4 Role of Automated Mechanisms

As I discussed in the introduction of this section, there is a trade-off between efficiency and costs of content moderation. It is possible that moderation systems could largely prevent spam, harassment and other forms of abuse, even on large communities, if enough expert human moderators are available to carefully review each post, but that may drive up the costs of moderation to unacceptable levels. One potential solution to minimize such costs is to automate content regulation. Recognizing this, some scholars have begun developing automated (often, machine learning based) solutions to automate certain aspects of content regulation (Gollatz, Beer, and Katzenbach, 2018). For example, researchers have proposed computational approaches to identify hate speech (Chandrasekharan et al., 2017a), pornography (Singh, Bansal, and Sofat, 2016) and pro-eating disorder content (Chancellor et al., 2017). Wulczyn et al. created a machine learning classifier trained on human-annotated data to identify personal attacks in online discussions on Wikipedia (Wulczyn, Thain, and Dixon, 2017). Park et al. designed a visual analytic tool, CommentIQ, that can help moderators select high-quality comments on online news sites at scale (Park et al., 2016).

Alongside the growing scholarly interest in automated moderation, many platforms are also increasingly deploying tools that automate the process of identifying problematic material and taking appropriate moderation decisions (Madrigal, 2018). For example, The Washington Post uses ModBot, a software application that employs NLP and machine learning techniques, to automatically moderate user-comments on news articles (Jiang and

Han, 2019). Although not machine-learning based, Reddit Automoderator is an excellent example of automated systems that are widely embraced by online communities (Melendez, 2015; Renfro, 2016). There is optimism in the industry that AI and machine learning systems would eventually replace the thousands of human workers who are currently involved in making moderation decisions, either voluntarily or as paid workers (Madrigal, 2018). Such systems also have the potential to execute moderation in a much faster way than human moderators.

Despite the enthusiasm for automated moderation systems among social media platforms as well as researchers, such systems face many challenges. Critics have argued that the currently available AI technologies are not good at understanding the context of a given post, user or community (Madrigal, 2018), so they may end up resulting in many false positives[1], that is, posts that are not problematic to the community get removed. Worse still, Blackwell et al. (2017) found that using automated approaches to identify abusive language can result in situations where the concerns of only the dominant groups are emphasized and existing structural inequalities are exacerbated. These systems are also vulnerable to the same challenges that human moderators face – many moderation decisions are complex, subtle and contested, and different users may feel differently about the appropriateness of the same content (Gillespie, 2018a).

Moderators of many Reddit communities configure and manage an automated moderation tool called Automod (Morris, 2015). While this solution is not machine learning (ML)-based, it still automates moderation for a large number of posts and comments. In Chapter 5, I show how moderators use this tool and the benefits and limitations of using this tool. I studied content moderation on Reddit as a sociotechnical system (Mumford, 2000). This allowed me to attend to the complexities generated by the coupling between

[1] I call a post a true positive if the moderation system removes that post and the moderators consider that post removal appropriate. Although it might seem counter-intuitive to use the term ‘true positive’ to denote a correctly *removed* post, it highlights the focus of moderation system on removal of inappropriate content. Besides, this is in line with my moderator participants’ use of the term ‘true positive’ to refer to correct removals.

technical and social components of this system (Opazo, 2010).

My work on automated moderation tools is related to Geiger and Ribes' research on the use of software tools to enforce policies and standards on Wikipedia (Geiger and Ribes, 2010). Geiger and Ribes showed that bots on Wikipedia automatically revert edits to its pages based on "criteria such as obscenity, patent nonsense, mass removal of content, and various metrics regarding the user who made the edit." They also found that assisted editing programs show new edits to Wikipedia editors in queues and allow them to quickly perform actions such as reverting edits or leaving a warning for the offending editor (Geiger and Ribes, 2010). My work adds to this research by describing the coordinated actions of moderators and automated tools in the context of Reddit. Although there are substantial differences in design mechanisms, outputs of content regulation, and functioning of software tools between Wikipedia and Reddit, Geiger and Ribes' work provides an excellent point of comparison to my research on how human and non-human work can be combined to facilitate content moderation.

Additionally, in Chapter 7, I will highlight the role that automated tools play in providing removal explanations on Reddit. I also scrutinize whether explanations provided by automated tools impact user behaviors differently than explanations provided by human moderators.

2.1.5 Fairness and Transparency in Moderation Systems

Social media platforms play a decisive role in promoting or constraining civil liberties. They make day-to-day judgments about which content is allowed and which is removed, and intervene in public disputes over intellectual property and controversial speech (DeNardis and Hackl, 2015). Indeed, how platforms make these decisions has important consequences for the communication rights of citizens and the shaping of our public discourse (Gillespie, 2018a). Therefore, the question of whether these platforms enact fairness in how they moderate user posts is important to consider.

Prior research has suggested a number of different frameworks in which platforms can ground their policy decisions, each with its own merits and challenges (Gorwa, 2019; Gill, Redeker, and Gasser, 2015a; Suzor, 2018; Suzor, Van Geelen, and Myers West, 2018; Suzor et al., 2019). Each of these frameworks emphasize different sets of normative principles, ranging from rights-based legal approaches (Gill, Redeker, and Gasser, 2015a) and American civil rights law (Citron, 2009) to principles of social justice (Gorwa, 2019; Suzor, Van Geelen, and Myers West, 2018; Taylor, 2017). Another framework that has been proposed for platform governance is the 'fairness, accountability, transparency and ethics' (FATE) model, that has recently seen a lot of interest in the HCI community, especially as it applies to algorithmic systems (Chancellor et al., 2019; Diakopoulos et al., 2017; Gorwa, 2019). Although this model has faced some critiques, such as its slowness to fully incorporate the lessons of intersectionality (Hoffmann, 2019) and its tendency to overstate the power of technology (Peña Gangadharan and Niklas, 2019), I draw from this model as a starting point, and focus on the principles of fairness and transparency in my work (Chapters 6 and 7) because this allows me to engage with other relevant HCI literature that concerns fairness and transparency in sociotechnical systems (Eslami et al., 2019; Jhaver, Karpfen, and Antin, 2018; Rader, Cotter, and Cho, 2018).

Fairness in Moderation Systems

While answering the question of whether the content moderation on a site is fair is a non-trivial exercise, it is notable that in recent years, social media platforms have often been criticized for how they moderate their content. Media sources have frequently reported how platforms fail to remove disturbing content (Citron, 2014; Nagle, 2017; Romano, 2017), block content of civic importance (Grenoble, 2013; Kelion, 2013; Scott and Isaac, 2016), promote content that is related to conspiracy theories (Allcott and Gentzkow, 2017; Uscinski, DeWitt, and Atkinson, 2018), and show political biases in content curation (Bozdag, 2013; Carlson, 2018). Some scholars have also begun raising questions about the appropriate limits on the private exercise of power by social media platforms (Gill, Redeker, and Gasser, 2015b; Suzor, Van Geelen, and Myers West, 2018), highlighting the human rights values that must inform the evaluation of fairness in content moderation. Such values include “freedom of expression”, “due process” and “transparency and openness” (Suzor, Van Geelen, and Myers West, 2018).

I add to this conversation on fairness in content moderation by providing empirical insights into how end users perceive the concept of “fairness” in content moderation in Chapter 6. Understanding the perspectives of end-users is crucially important because they are the most important stakeholders in content moderation systems inasmuch as they are the ones who produce as well as consume content on these platforms as well as bring in the ad revenues to sustain them. In Chapter 6, I discuss my findings on how end-users’ perspectives of fairness in moderation decisions are affected by different elements of content moderation and the context of submissions.

This work is most closely related to the line of research that focuses on the perspectives of end-users of centralized moderation platforms (Grimmelmann, 2015) like Facebook and Twitter. For example, in a recent study, West analyzed users whose posts are removed on Facebook, Twitter and Instagram, and surfaced these users’ folk theories of how content moderation works (West, 2018). Chapter 6 complements this prior research on centralized moderation platforms by highlighting the impact of content moderation on users in the distributed moderation system of Reddit. I expect that the multi-community environment of Reddit, along with its locally crafted guidelines and volunteer-driven moderation, makes moderation on Reddit a much different experience from moderation on other platforms such as Facebook and Twitter. My research on fairness in moderation attempts to highlight the concerns that arise in the distributed moderation system of Reddit. I use the user-centered perspectives I obtain in my findings to suggest guidelines on designing moderation tools and strategies that may improve the health of distributed online communities like Reddit.

Moderators on many sites use explicit community guidelines to make the community norms more visible, especially to newcomers (Fiesler et al., 2018). However, how these guidelines affect user attitudes and behaviors remains unclear. Kiesler et al. hypothesize that while explicit guidelines may help users understand what is acceptable to post on a community, they may also discourage user contributions if the users feel stifled (Kiesler, Kraut, and Resnick, 2012). Building upon their theoretical work, I provide empirical insights into how the presence of community guidelines and the perceptions of users about these guidelines shape their attitudes about the community.

Chapter 6 also specifically extends prior work on community-created rules on Reddit. Fiesler et al. (2018) recently presented a description of the rules ecosystem across Reddit, highlighting that the activity on Reddit is guided by multiple layers of rules. First, there exists a user agreement and content policy similar to the terms and conditions of many websites. Second, a set of established rules defined by Reddit users, called Rediquette, guide site-wide behavior. Finally, many subreddits also have their own set of rules that exist alongside site-wide policy and lay out expectations about content posted on the community (Fiesler et al., 2018). Thus, Reddit users operate in an ecosystem of governance where even the explicit guidelines for expected behavior, let alone the implicit norms, derive from multiple sources. Prior research has shown that negotiating multiple sources of rules can result in confusion among end-users (Fiesler, Feuston, and Bruckman, 2015). I explore in Chapter 6 how Reddit users negotiate posting in multiple communities, each having its own set of rules, and the challenges they face in the process.

This survey study (alongside Chapters 3 and 4) also contribute to the growing body of research on understanding bad actors online (Blackwell et al., 2018b; Blackwell et al., 2018c; Coleman, 2014a; Phillips, 2015a). Coleman (2014a) and Phillips (2015a) both conducted deep ethnographic investigations to understand the subculture of internet trolls. I add to their research by surfacing the perspectives of bad actors on Reddit and discussing various factors that motivate them to post. I also explore how the difficulties in identifying bad actors complicate the process of allotting moderation resources to nurturing sincere users.

Folk Theories of Sociotechnical Systems

DeVito et al. define folk theories as "intuitive, informal theories that individuals develop to explain the outcomes, effects, or consequences of technological systems, which guide reactions to and behavior towards said systems" (DeVito, Gergle, and Birnholtz, 2017). In recent years, HCI and CSCW researchers have begun exploring how users of sociotechnical systems develop folk theories about their operations and the ways in which such folk theories affect users' interactions with these systems (DeVito, Gergle, and Birnholtz, 2017; DeVito et al., 2018; Eslami et al., 2016; French and Hancock, 2017; Jhaver, Karpfen, and Antin, 2018).

DeVito et al. analyzed how Facebook users seek and integrate different information sources to form folk theories of algorithmic social media feeds and how these folk theories interplay with users' self-presentation goals (DeVito et al., 2018). In Chapter 6, I discuss the need for similar efforts to understand folk theory formations in content moderation systems.

Jhaver et al. investigated the folk theories developed by Airbnb hosts about the operation of Airbnb search algorithm and found that hosts' beliefs in some of these theories created anxiety among them, often forcing them to engage in wasteful activities as part of their coping strategies (Jhaver, Karpfen, and Antin, 2018). Similar to this, I found folk theories in my survey data (Chapter 6) that caused anxieties and frustrations among end-users in the context of Reddit moderation. Eslami et al. studied the folk theories of Facebook News Feed curation algorithm and concluded that implementing structured transparency or "seams" into the design of these systems may help improve human-algorithm interaction and benefit human agency in complex systems (Eslami et al., 2016). I add to this literature by exploring the interplay between folk theories and transparency in the domain of content moderation. I highlight the folk theories of content moderation that Reddit users develop in order to make sense of their content removals, and discuss how these theories may inform governance practices.

Transparency in Moderation Systems

Cornelia Moser defines transparency as opening up "the working procedures not immediately visible to those not directly involved in order to demonstrate the good working of an institution" (Moser, 2001). Although transparency is not a new idea in governance, it has recently drawn a new surge of interest because of the transforming powers of digital technologies (Fung, Graham, and Weil, 2007). Internet and mobile technologies have reduced the information asymmetries between organizations and customers by facilitating instant dissemination of knowledge. Consequently, end-users increasingly expect to be very well-informed (Granados and Gupta, 2013). Many media articles and non-profit organizations have advocated the virtues of transparency, and connected it to trust, corporate social responsibility and ethics (Rawlins, 2008). As a result, many organizations are increasingly declaring themselves as transparent in order to gain the trust of their customers.

While transparency can be seen as a means to ensure social accountability (Breton, 2007), the process of adopting a transparency strategy is far from trivial for organizations. They need to account for the "complex dependencies, trade-offs, and indirect effects" of disclosing each informational element to their customers and competitors (Granados and Gupta, 2013). In the context of content moderation systems, social media platforms have to consider the effects of transparency not just on individual users but also on news media that are increasingly being critical of moderation processes (Scott and Isaac, 2016).

Recently, HCI and communications researchers have begun reflecting on the importance of transparency in content moderation (West, 2018; Seering et al., 2019b; Suzor, Van Geelen, and Myers West, 2018; Suzor et al., 2019). For example, West noted that moderation systems have the opportunity to serve an educational, rather than a punitive, role by providing moderated users an understanding of why their content was removed (West, 2018). Suzor et al. have called for a range of digital platforms to issue transparency reports about how they enforce their terms of service (Suzor, Van Geelen, and Myers West, 2018). However, there still exists a gap in our understanding of how transparency in distributed moderation systems like Reddit affect user attitudes and behaviors.

I begin to fill this gap by focusing on explanations for content removals, an important aspect of transparency in moderation on social media platforms. Explanations are difficult to design effectively (Bunt, McGrenere, and Conati, 2007; Gregor and Benbasat, 1999; Herlocker, Konstan, and Riedl, 2000), and they require time and effort on the behalf of both moderators who provide them as well as users who consume them. Thus, it is unclear whether they are effective and worth implementing. In Chapters 6 and 7, I examine the effectiveness of explanations from the perspective of end-users. First, in Chapter 6, I focus on how users perceive such explanations, how they feel about the lack of any explanations, and the ways in which the content, modes and sources of explanations affect user perceptions. Following this later on, in Chapter 7, I empirically examine the ways in which providing explanations for content removals affect the future behavior of users. I see these studies as some of the first steps in understanding transparency and explanations in content moderation. These chapters provide insights into how we can design transparent moderation systems so as to encourage active and healthy participation.

2.1.6 Moderation Research on Specific Sites: Reddit and Twitter

My thesis has focused on content moderation of two major social media platforms - Reddit and Twitter. I concentrated on these two platforms because both of them are enormously popular sites, boasting millions of daily active users, allowing me to problematize the moderation systems that are seemingly successful. These platforms suffer from the challenges of scale that make the process of content moderation significantly more difficult, so focusing on them allowed me to explore how high-volume content necessitates alternative moderation strategies such as the use of automated and third-party moderation tools. Besides, these two platforms share much in common with other past and present moderation systems that have been studied in prior research, so focusing on them opened up opportunities for me to engage closely with this literature. Indeed, Reddit has a decentralized moderation system with many structural and procedural similarities to moderation on other popular distributed sites like Facebook Groups, Wikipedia and Discord. Similarly, Twitter has a centralized moderation system with many of its features similar to the moderation systems on other centrally operated sites like Facebook and Instagram. Therefore, although my thesis constitutes studies about content moderation on Twitter and Reddit only, the insights drawn from this work are easily transferable to a number of other social media platforms.

Next, I briefly describe the moderation infrastructure on Twitter and Reddit sites. These descriptions provide context on the layout and affordances of the two platforms. I encourage readers unfamiliar with Twitter and/or Reddit to attend to these subsections because the concepts and dynamics I discuss in the following chapters are seated in the design mechanisms of these platforms.

Reddit Moderation

Reddit is composed of thousands of small and large communities called subreddits where users can post submissions or comment on others' submissions. Activity on Reddit is guided by a user agreement[2] and content policy[3] similar to the terms and conditions of many websites and a set of established rules defined by Reddit guidelines called Reddiquette[4] (Fiesler et al., 2018). Each subreddit also has its own set of rules that exist alongside site-wide policy and lay out what content is acceptable and what is not acceptable on that subreddit. These rules are typically found in sidebars of the subreddit (Matias, 2016b). They have higher salience than Reddiquette for most users (**Kiesler2012**). Many subreddits have a separate set of rules for submissions and comments.

[2] https://www.reddit.com/help/useragreement/

[3] https://www.reddit.com/help/contentpolicy/

[4] https://www.reddit.com/wiki/reddiquette

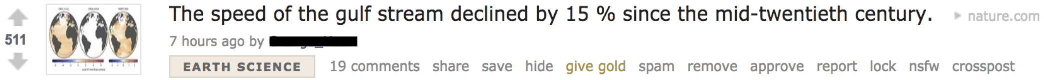
Reddit moderators are volunteer Reddit users who take on the responsibility of maintaining their communities by participating in a variety of tasks. These tasks include (1) coordinating with one another to determine policies and policy changes that guide moderation decisions, (2) checking submissions, threads and content flagged by users for rule violations, (3) replying to user inquiries and complaints [5], (4) recruiting new moderators, (5) inviting high-profile individuals to conduct AMA (Ask Me Anything) sessions (Moreau, 2017), (6) creating bots (Long et al., 2017) or editing Automod rules (described below in this section) to help automate moderation tasks, and (7) improving the design of the subreddit using CSS tools. Moderators usually prefer to focus primarily on a few of these task categories depending on their interests, skills, prior moderation experience, level of access[6], influence among the moderators and the requirements of the subreddit.

One automated solution to content regulation is "Automoderator" (or "Automod"), which first became popular as a third-party bot but was later incorporated into Reddit and offered to all the moderators. Many moderators use Automod to help improve their work efficiency. This solution uses a filtering approach where moderators codify phrases that usually occur in undesirable posts as regular expressions[7] (also called 'regex') into a wiki which they regularly update. Automod then scans each posted material for the presence of the coded phrases and removes the material if it contains any such phrase. In Chapter 5, I investigate the use of Automod tool as a component of the sociotechnical system of content regulation. I also briefly discuss how other bots are used to improve the efficiency of moderation tasks. My findings in this chapter highlight how the currently employed human-machine collaborations help manage content quality on Reddit.

[5] Users can reach out to moderators using ModMail, a personal messaging tool that forwards inquiries to all moderators of a subreddit.

[6] Moderators can be given different levels of access on a subreddit depending on their roles. Different binary flags can be set to provide different permissions. For example, 'access' flag allows moderators to manage the lists of approved submitters and banned users, and 'posts' flag allows moderators to approve or remove content. Only moderators with 'full permissions' can change the permission levels for other moderators.

[7] Regular expressions are special text strings for describing specific search patterns. They are used to search a volume of text for a group of words that match the given patterns (Thompson, 1968).



The speed of the gulf stream declined by 15 % since the mid-twentieth century. ▸ nature.com

7 hours ago by [redacted]

Figure 2.1: Moderators' view of a submission post. This interface differs from a regular user's interface because of the presence of additional links for 'spam', 'remove', 'approve', 'lock' and 'nsfw'. Moderators can use these links to take corresponding actions on the post. For example, clicking 'lock' prevents the submission post from receiving any new comments.

Reddit provides its moderators with alternate interfaces that contain tools that are not accessible to regular users of the subreddit (see Figure 2.1). These tools can be used to perform a variety of moderation actions. For example, for each submission, moderators can remove the submission thread, lock the thread from receiving any new comments, or remove any comment on the submission thread. Each subreddit also has its own private moderation log, a tool that allows moderators to view which posts and comments have been removed, at what time, and by which moderator. Although Reddit provides these moderation tools to all moderators by default, many moderators find these tools inefficient and cumbersome to use. This has motivated the creation of third-party front-end tools that help moderators regulate their communities more efficiently. I discuss how such tools and back-end bots are built and shared between different communities in my findings (Section 5.3.1).

Twitter Moderation

To the best of my knowledge, my work on Twitter moderation, as described in Chapter 4, is the only major empirical research on Twitter moderation that has been conducted so far. By contrast, quite a few researchers have focused on understanding content moderation on Reddit (Chandrasekharan et al., 2017a; Chandrasekharan et al., 2018; Fiesler et al., 2018; Matias, 2016c; McGillicuddy, Bernard, and Cranefield, 2016). Matias studied the ways that Reddit moderators negotiate how they are viewed by platform operators, community participants and other moderators in various aspects of their work (Matias, 2016c).

Matias also showed how Reddit moderators respond to users' complaints of censorship (Matias, 2016c). Fiesler et al. (2018) conducted a mixed-methods study of 100,000 subreddits and contributed a comprehensive description of the type of rules that are enacted across Reddit. Kiene et al. studied the role of moderators in welcoming newcomers to rapidly growing subreddits (Kiene, Monroy-Hernández, and Hill, 2016). McGillicuddy et al. studied the political and moral struggles that Reddit moderators face in their work (McGillicuddy, Bernard, and Cranefield, 2016). They emphasized the moderators' awareness of their power over the community. McGillicuddy et al. also showed how moderators codify norms to dictate behavior and hire new moderators (McGillicuddy, Bernard, and Cranefield, 2016). I contribute to this growing body of research on Reddit moderation by investigating the challenges of moderating against online harassment in Chapter 3, highlighting the role of technology in assisting the work of content moderators in Chapter 5, and exploring the concepts of fairness and transparency in content moderation in Chapters 6 and 7. I distinguish my work from prior research on social dynamics of Reddit moderators (e.g., (Matias, 2016c; Matias, 2016a; McGillicuddy, Bernard, and Cranefield, 2016)) by bringing to scrutiny the moderator interactions that are related to the use of Automod and the provision of removal explanations. I also build upon my findings to discuss the design implications for building mixed-initiative, inclusive, fair and transparent content moderation systems.

2.2 Online Harassment

In recent years, the problem of online harassment has reached alarming proportions, and is afflicting an ever-increasing number of individuals. According to a 2016 Data & Society Research Institute report based on a nationally representative survey, 47% of internet users have experienced online harassment (Lenhart et al., 2016).

Like face-to-face harassment, online harassment has a deeply negative impact on its recipients. They may experience significant emotional problems such as anxiety and depression and, in extreme situations, may even commit suicide (Ashktorab and Vitak, 2016). A 2017 Pew Research study found that 13% of US adults experienced mental or emotional stress as a result of online harassment, another 8% of adults indicated that they had problems with friends or family because of online harassment, and 7% said that these experiences caused damage to their reputation (Duggan, 2017). Lenhart et al. (2016) pointed out that 27% of internet users self-censor their online postings out of fear of online harassment. Furthermore, 43% of recipients of online abuse had to change their contact information to escape their abuse.

Young people, minorities and women are particularly vulnerable to such abuse (Duggan, 2014; Duggan, 2017; Lenhart et al., 2016). In a national study of middle and high school students, 60 percent of lesbian, gay, bisexual, and transgender (LGBT) youth reported being harassed based on their sexual identity and 56 percent of them felt depressed as a result of being cyberbullied (Cooper and Blumenfeld, 2012).

One reason why combating online harassment is challenging is that it is often difficult to reach a consensus on which action crosses a line and which doesn't. Although online harassment has attracted a lot of media attention and research interest in recent years, there is no standard agreed-upon definition of what online harassment entails. This makes it difficult for social media platforms to battle abusive behaviors because they don't want to be seen as censoring what people can say online. A discussion of how to efficiently address online harassment requires that we start having conversations about what exactly it is.

Some researchers have attempted to conceptualize online harassment. Ybarra and Mitchell (2004) define it as "an intentional and overt act of aggression toward another person online" . Lwin, *et al.* define online harassment as "rude, threatening or offensive content directed at others by friends or strangers, through the use of information communications technology" (Lwin, Li, and Ang, 2012). However, what should be considered "offensive content" or an "act of aggression" can be subjective.

Lenhart et al. (2016) argue that "online harassment is defined less by the specific behavior than its intended effect on and the way it is experienced by its target". Many users who are accused of perpetrating harassment, however, complain that simple disagreement on their part is often portrayed as harassment by other users. Everyone agrees that posts of death threats and rape threats are abusive and should be regulated. But beyond such posts, where do we draw the line? How do we distinguish someone deliberately trying to harm another user from someone passionately disagreeing with that user? I explore these questions in Chapters 3 and 4.

Several researchers have highlighted the unique characteristics and challenges of online harassment. In their review of research on solicitation, harassment and cyberbullying, Schrock and boyd found that there is often an overlap between harassers and victims of harassment (Schrock and Boyd, 2011). They also noted that "offending parties are frequently anonymous and include both adults and youths." Danah boyd has identified four properties of SNSs that fundamentally alter the social dynamics of harassment and amplify its effects online: persistence, searchability, replicability, and invisible audiences (danah, 2008). The internet exacerbates the harassed users' injuries by extending the life of destructive posts. Search engines index content on the web, and harassers can put indexed abusive posts to malicious use years after they first appear (Citron, 2014). Furthermore, by increasing the visibility of content, SNSs enable harassers to call for others to engage in abusive behaviors (Vitak et al., 2017).

I build on this prior work in my thesis to understand more comprehensively and rigorously the different aspects of online harassment, its types, and its impact on those who are harassed.

2.3 Freedom of Speech on Internet Platforms

The First Amendment to the United States Constitution prohibits the government from censoring what citizens can say: "Congress shall make no law ... abridging the freedom of speech, or of the press." The notions of free speech in the U.S. are unique, even among democratic countries. Many Western democracies like The United Kingdom, France, Germany and Netherlands have laws against hate speech or speech that incites racial hatred. In contrast, the U.S. protects free speech to the extent that it even protects racially and religiously offensive material (Oetheimer, 2009; Zoller, 2009). For example, most of Western Europe has banned Holocaust denial, but there are no laws against it in the U.S.

Many contemporary First Amendment scholars generally believe that the risks of limiting expression always outweigh the risks of giving free rein to speech, and they advocate protecting all utterances and publications without discrimination (Canavan, 1984). They view freedom of expression as an absolute, overriding end in itself and warn of the dangers of the "slippery slope" — a current acceptable change to the status quo regarding speech can lead to some intolerable future limitations on speech if speech prohibition is introduced, they argue.

Some scholars have challenged this absolutist interpretation of the First Amendment. For example, Francis Canavan asserts that the purpose of the First Amendment is to protect and facilitate the achievement of rational ends by communication among ordinarily intelligent people (Canavan, 1984). He argues that when speech does not contribute to this purpose, it should not be protected.

The ever-increasing use of internet for public conversations further complicates these issues. As more and more of all discourse moves to online platforms like Facebook, Twitter and Reddit, citizens cede control over what may be said to corporate policies and procedures that are largely unregulated and for which there is no recourse. As I will discuss in further detail below, many scholars have expressed concerns about this privatization of internet content control (DeNardis, 2012; DeNardis and Hackl, 2016; Nunziato, 2005; Schesser, 2006). The internet also creates new jurisdictional challenges for regulating content, because users from different countries with different laws on what speech is acceptable can post on the same forums.

In the early days of the Internet, many academics and entrepreneurs of Silicon Valley shared a vision of freedom on the internet. They envisioned the internet as a forum for users to connect with one another without regard to race, gender, age or geography. They opposed legislation that would increase government regulation of the web and expected that the web would be adequately governed by the users themselves (Wortham, 2017).

However, this optimistic vision underestimated what would happen as the internet grew. Unfortunately, with the growing opportunities for users to connect with one another online through social network sites, the internet has also attracted a lot of disturbing behaviors. Such behaviors have been found to be pervasive and difficult to regulate on many sites. One reason for this failure is that the founding values of freedom on the internet are so ingrained that many users and critics dislike regulation of any content on the web. Many internet companies also embrace a libertarian view and try to avoid regulation in technology (Wortham, 2017).

A number of scholars have urged caution about the consequences of allowing extreme forms of speech. For example, Nancy Kim argues that applying the First Amendment analysis to the free speech versus online harassment debate without recognizing the ways in which online communication differs from off-line communication fails to address many of the harms of online harassment (Kim, 2009). When harassment is conducted online, it can have more serious consequences because it is easier for digital information to spread faster and more widely. Danielle Citron argues that an absolutist devotion to free speech "needs to be viewed in light of the important interests that online harassment jeopardizes" (Citron, 2014). She writes that many harassers silence their victims and "we need to account for the full breadth of expression imperiled" when we evaluate the risks to expression that exist in our efforts to regulate online abuse. Even though many Americans consider free speech a universal and self-evident right, free speech is both a cultural norm and a legal construct, and its interpretations vary widely across different countries, thereby creating new challenges for content regulation online.

Some critics argue that although hate speech is legal in the U.S., internet platforms do not need to allow it. Laws about freedom of speech in the U.S. control what the government can do — not private individuals. A corporate platform is more akin to a private party or club and it is legal to set rules for what may and may not be said there. However, other scholars have criticized the privatization of internet content control and lament the absence of places on the internet where free speech is constitutionally protected (Nunziato, 2005; Schesser, 2006) . DeNardis has argued that since globalization and technological change have reduced the capacity of sovereign nation states to control information flows, internet governance has now become the central front of freedom of expression (DeNardis, 2012; DeNardis and Hackl, 2016). In the same vein, Nunziato has noted the compelling private interests to provide public online spaces and advocated for legislatures to faithfully translate First Amendment values in cyberspace in order to make them meaningful in the technological age (Nunziato, 2005).

The overview of free speech I present above is brief and far from exhaustive. However, it starts to reveal some of the challenges of regulating abusive behaviors and distinguishing disagreements from online harassment. I argue that online activities that are intentionally aggressive towards other individuals and threaten them should be considered as instances of online harassment. However, it is challenging to regulate actions which may be considered impolite and offensive by some individuals, but which are civil and conducive to fostering democratic goals.

I use this perspective in Chapter 3 to present an emic account of the subreddit Kotaku in Action. As I will show, members see their group and their practices quite differently from their typical portrayal in both the popular press and academic literature. In seeing things from KiA members' point of view, I develop more broadly relevant insights into the boundary between freedom of expression and harassment.

I also use the lens of difficulties in distinguishing appropriate and inappropriate content to guide all the other chapters in my thesis. For example, I explore how the use of automated mechanisms in content moderation affect the processes that make such distinctions in Chapter 5. I discuss in Chapter 4 how the users' disillusion with deficiencies in current processes motivate them to create and/or use third-party tools like Twitter blocklists, and how the use of these tools, which have no central oversight, in turn, create new types of censorship challenges. Finally, in chapters 6 and 7, I investigate the different ways in which explanations about decision making in moderation processes are communicated to the users, how more often than not,such decisions are made opaquely and without any explanations, and the ways in which such communications (or their absence) impact users and moderators.

CHAPTER 3
CONCEPTUALIZING ONLINE HARASSMENT: THE CASE OF KOTAKU IN ACTION

In this chapter, I use mixed methods to study a controversial Internet site: The Kotaku in Action (KiA) subreddit. Members of KiA are part of GamerGate, a distributed social movement. I present an emic account of what takes place on KiA: who are they, what are their goals and beliefs, and what rules do they follow. Members of GamerGate in general and KiA in particular have often been accused of harassment. However, KiA site policies explicitly prohibit such behavior, and members insist that they have been falsely accused. Underlying the controversy over whether KiA supports harassment is a complex disagreement about what "harassment" is, and where to draw the line between freedom of expression and censorship. I propose a model that characterizes perceptions of controversial speech, dividing it into four categories: criticism, insult, public shaming, and harassment. I also discuss design solutions that address the challenges of moderating harassment without impinging on free speech, and communicating across different ideologies[1].

3.1 Introduction

In a 2015 podcast on the radio show "This American Life," writer Lindy West interviews a man who viciously harassed her over the Internet (West (Speaker), 2015). In the process, her harasser comes to recognize her humanity and apologizes for his behavior. In return, West comes to understand her harasser as a person, and the personal challenges that explain (though don't excuse) his behavior. Inspired by West's experiences, I decided to try to understand online harassment in a more nuanced way by talking directly to harassers.

[1] Findings from this study were published in 2018 in the First Monday journal (Jhaver, Chan, and Bruckman, 2018). Larry Chan assisted me and Amy Bruckman guided me on this work.

Specifically, I chose to study Kotaku in Action (KiA), a discussion forum for members of GamerGate. GamerGate is an online social movement portrayed in the popular press as a misogynistic hate group. A Washington Post article described GamerGate as "the freewheeling catastrophe/social movement/misdirected lynchmob that has, since August, trapped wide swaths of the Internet in its clutches, [and] has still — inexplicably! — not burned itself out." (Dewey, 2014). What is behind that portrayal? In this research, I wanted to understand: Who are these people and how do they see their group and their activities? How do the features of their sociotechnical system and the popular perceptions of their activities influence their behavior? What can we learn about controversial speech and harassment by considering their perspectives?

As a researcher, I value *listening*. Individuals usually have more complex views than stereotypes predict. It is valuable to listen to the point of view of *anyone*, and listening does not imply either accepting or rejecting that person's view of the world. Inspired by Coleman's studies of the group Anonymous (Coleman, 2014b) and Phillips' work on trolls (Phillips, 2015b), I see value in trying to understand how members of KiA understand their commitments and practices. *Nothing in this chapter should be construed to either support or attack either side of the GamerGate controversy.* The point of this work is to use this rich context to develop a more nuanced understanding of the complex boundary between free speech and harassment.

I begin with a brief description of GamerGate and Kotaku inAction community so as to clarify the context in which this study was conducted. Next, I present my methods of data collection and analysis. I divide my findings into two sections: a portrait of KiA (its members and activities), and members' views on free speech and harassment. Finally, I discuss broader implications of this research for theory and design.

3.2 Gamergate and Kotaku in Action

3.2.1 GamerGate

GamerGate is a distributed social movement that emerged in August 2014 (Kain, 2014). It began with a series of controversial events surrounding game developer Zoe Quinn, who was harassed as a result (Jason, 2015). This sparked a broader movement which initially focused on "ethics in game journalism," but quickly expanded to address broader issues of censorship, negative stereotyping of nerd culture, "social justice warrior" (SJW)[2] ideology, and perceived excesses of political correctness (Glasgow, 2015). Mortensen provides a detailed account of the progression of events that helped GamerGate gain popular attention (Mortensen, 2016).

Opponents of GamerGate have experienced doxing (revealing someone's personal information online), death threats, rape threats, and SWATing (tricking the police into raiding someone's home). Many who support GamerGate disavow these tactics (Glasgow, 2015), and insist that the perpetrators do not represent them. GamerGate does not have a defined membership or official leaders, so it is difficult to state whether "GamerGate" committed any particular act. Moreover, many GamerGate supporters claim that they also experienced doxing and harassment.

Like many contemporary online phenomena, GamerGate is not something that happens on one online site, but on a collection of sites with complex interactions across them (Gonzales, Fiesler, and Bruckman, 2015). Supporters use the hashtag #GamerGate on Twitter, and websites like Reddit, 8chan, Voat and Tumblr for communication and collaboration (Wiki, 2016).

Although I initially set out to study GamerGate more generally, it rapidly became apparent to me that this was an impossibly huge task. To focus my efforts, I chose one online

[2] Among GamerGate supporters, "social justice warrior" or SJW is a pejorative term for someone who, they claim, repeatedly makes shallow arguments about social justice for the purpose of raising their own personal reputation (Know Your Meme, 2016)

community dedicated to the discussion of GamerGate issues: the Kotaku in Action (KiA) subreddit (Kotaku In Action, 2016).

3.2.2 Kotaku in Action

KiA describes itself as “the main hub for GamerGate discussion on Reddit.” It is titled “Kotaku in Action” because at the time of its creation, it was dedicated to satirize Kotaku (a news and opinion site about games) for its alleged unethical journalistic practices. The sidebar of KiA declares its mission as: “KotakuInAction is a platform for open discussion of the issues where gaming, nerd culture, the Internet, and media collide.”

As of July 7, 2019, KiA has 113,046 subscribers, and hundreds of active users at any given time. Its discussion board continues to remain active with dozens of new submissions every day.

KiA is just one of many sites for discussions on GamerGate. In fact, Reddit itself hosts a number of other subreddits for discussions related to GamerGate like “SocialJusticeInAction” and “GGDiscussion.” A popular multi-reddit (a Reddit feature that enables combining and subscribing to several subreddits together) called “KiA HUB” collates submissions from ten such subreddits. There also exists subreddits like “r/GamerGhazi,” that are devoted to anti-GamerGate discussions.

As Treré points out, “restricting the focus to only one of the many online technological manifestations of social movements risks overlooking important aspects such as the role and evolution of different platforms within a movement” (Treré, 2012). Therefore, I must note here that to fully understand the dynamics of GamerGate, future research must also consider Twitter as well as comparatively obscure sites like 8chan and Voat used by GamerGate supporters and opponents.

KiA is generally one of the mildest GamerGate forums, with less controversial speech than discussions of the topic on other sites like Twitter, 8chan or Voat. Although the popular press portrays GamerGate as a movement of misogynist Internet trolls (Allcott and

Gentzkow, 2017), I found that KiA members do not view themselves as such. Values my participants embrace include a strong support for freedom of speech, the view that political correctness has gone too far in our society, the idea that white men are discriminated against in today's society, and a belief that the quality of journalism is in decline and the mainstream press too often blindly follows the values of PC (politically correct) culture.

3.3 Methods

I begin this section by briefly discussing how I created rapport with the KiA community. Jennifer A. Rode writes that "discussions of rapport, even the cultural bumbling of getting it wrong, is critical to the ethnographic enterprise" (Rode, 2011). In particular, KiA was suspicious of outsiders studying it because it felt betrayed by previous occasions of journalists misrepresenting the community after interacting with its users. I hope that this discussion of rapport building contributes to the reader's understanding of the nature of my ethnographic encounters and my findings. I follow this by a description of my methods, participants and analysis.

In a graduate class on online communities taught by my advisor Amy Bruckman, students complete a qualitative study of an online site using a combination of participant observation and interviewing. I began studying Kotaku in Action as part of a project team of three students. Not long after we began to request interviews on the site (following ethical guidelines (Bruckman, 2006)), this message was posted on KIA:

> "Dear KiA, if you are contacted by /u/gatech01[3] for this project, please be aware that this [is] indeed a trap, because the person doing the data collection and interpretation is intrinsically ideologically opposed to everything that this sub stands for."

In response to this, Amy replied, and volunteered (in Reddit tradition) to do an "AMA" ("ask me anything" discussion thread) with KiA users. The AMA took place on Feb 27,

[3] Reddit members use /u/username to identify a Reddit user by his/her username.

2016 and includes 262 comments (Bruckman (submitter), 2016). This is when the community began to tolerate our presence, and started providing us valuable information. Following the AMA, many KiA members volunteered to speak with the research team.

Our study was approved by the Institutional Review Board (IRB) of the Georgia Institute of Technology. In all, we conducted thirteen semi-structured interviews with KiA users. All the interviews were conducted in Spring 2016. We recruited participants through private messages on Reddit. Participation was voluntary, and no incentives were offered for participation.

The interviews generally lasted between 60 to 90 minutes. Participants were asked questions about how they came to use the subreddit, what motivated them to continue posting on the subreddit, and their views on online harassment and moderation on websites for discussions on GamerGate. We conducted interviews over the phone, on Skype, and through chat. Some participants were contacted for brief follow-up interviews, for further clarification. I also shared an early draft of the study with all the participants and they were given a chance to respond to it.

I read online postings of the participants, and compared their attitudes and actions to what they stated in their interviews. These were found to be largely consistent. However, because of the pseudonymous nature of Reddit platform, I constantly faced the possibility that my participants were lying in the interviews. Following Phillips' approach (Phillips, 2015b), I decided to note how the KiA users chose to present themselves to us, and deduce meaning from their (possibly choreographed) performance. Therefore, I present my findings as subjective perspectives and narratives of KiA rather than as objective facts.

Becker and Geer argue that any social group has a distinctive culture and a set of common understandings that find their expression in a language "whose nuances are peculiar to that group and fully understood only by its members" (Becker and Geer, 1957). They suggest to fieldworkers that "both care and imagination must be used in making sure of meanings, for the cultural esoterica of a group may hide behind ordinary language used in special ways." In studying KiA community, I paid particular attention to the terminology used by its members during my observations and interviews, and examined it as a function of their assumptions and purposes (Taylor and Bogdan, 1998). For instance, I have analyzed what terms like "political correctness" and "sealioning" mean to this community.

3.3.1 Participants

Eleven participants in the study reported ages between 24 and 37. Two participants did not share their age, but one of them mentioned that he is in his 20s. Twelve participants were male and one was female. Two of the participants were from Norway, and the rest from the US. Four of the participants chose not to share all the demographic information we asked for. The interviewees included one current and one past KiA moderator. Table 3.1 shows some demographic information about the participants. I use light disguise (Bruckman, 2002) in describing my findings. Therefore, although I have omitted sensitive details to protect the identity of my participants, some active KiA active members may be able to guess who is being discussed.

Table 3.1: Study Participants

ID	Age	Country	Occupation	Medium	Gender	Account Creation	Is Moderator?
P1	unknown	USA	Graphic Designer	chat	Male	Jun, 2015	No
P2	26	USA	Tech content manager	Skype	Male	May, 2015	No
P3	27	Norway	unknown	Skype	Male	Jan, 2015	No
P4	20's	USA	unknown	Skype	Male	Oct, 2009	No
P5	24	USA	unknown	Skype	Male	Mar, 2010	No
P6	24	USA	Grocery-store manager	Skype	Male	Oct, 2014	No
P7	25	USA	Chemist	Skype	Male	Nov, 2014	No

Table 3.1: Study Participants

P8	28	Norway	Computer Science student	chat	Male	Oct, 2013	Current moderator
P9	34	USA	Computer Programmer	chat	Male	Feb, 2013	No
P10	37	USA	Computer Technician	Phone	Male	May, 2014	No
P11	25	USA	Instructional aid	chat	Female	Feb, 2015	No
P12	24	USA	Student	chat	Male	May, 2012	Past moderator
P13	26	USA	Medical Professional	Skype	Male	Mar, 2011	No

Participants were self-selected—we spoke to members who were willing to be interviewed by a team of academics. Thus, my informants are likely some of the most moderate members of the group and are more engaged in talking about the topics at hand. They were eager to convince us and the world that they are reasonable people. I saw specific evidence for this fact by comparing what some of them said during their interviews to their often more raucous and rude public presences on Twitter. As a result, in my dataset, the most moderate people are on their best behavior. However, their views are nevertheless revealing, and speak to broader issues.

3.3.2 Analysis

I fully transcribed data from the interviews and undertook multiple readings of these data. I used Dedoose software (www.dedoose.com) to perform coding, which underwent multiple iterations. Next, I performed inductive thematic analysis on these data and identified relevant themes and sub-themes through observation and discussion (Braun and Clarke, 2006). I met regularly with all the co-authors of this study to discuss the codes and emerging themes, going back and forth between different categories and further scrutinizing data.

I conducted over 100 hours of participant observation of KiA in an attempt to systematically understand the dynamics of the community. This was carried out between January-May 2016, and included lurking, commenting, voting on content, and posting content on the subreddit. I supplemented our interview responses with field notes and qualitative analysis of posts. I also corresponded with the current KiA moderators, who answered my questions through private messages, and helped me improve my understanding of the community. Over the course of this research, my stances developed as a result of engagement with the community, and through my evolving understanding of its objectives and characteristics.

I employed a mixed-method design for this study. In addition to the traditional ethnographic-style methods for observing online communities discussed above, I also used quantitative methods for analyzing user behavior on KiA. I assembled a sample of 1000 random submissions on KiA. I used PRAW, a python package that provides access to Reddit's API (PRAW, 2016), to gather these submissions. KiA encourages users to tag their posts using flairs. These flairs indicate the topic of the submission (Table 3). I extracted the flairs used in the submissions I collected to find the most popular topics of discussion. I also analyzed the website domains of these submissions that were links to external websites. To find what other issues KiA users are interested in, I extracted the set of users who posted these submissions. Following this, I collected all the postings made by these users on Reddit, again using PRAW.

Exactly what GamerGate is about can be hard for outsiders to understand. In Appendix A, I present a case study that describes the kind of arguments that individuals on the opposite sides of GamerGate debate make. I encourage readers interested in understanding GamerGate to review this case study.

3.4 Findings

3.4.1 KiA Community

In this section, I describe members of Reddit who interact with the KiA subreddit. Many of these users identify as gamers. They comment, post or lurk on KiA's discussion board or moderate it. They use the subreddit to share news relevant to GamerGate, and engage in discussions with one another.

I begin with a discussion of the topics that KiA members are interested in. Next, I discuss the demographic diversity of KiA. This provides context for understanding the beliefs and goals of KiA members, which I examine in the next subsection. Next, I describe the practice of archiving employed by the community, a practice crucial to understanding the ethos of the community. Finally, I analyze the conventions and policies that guide the activity of KiA users. This analysis helps explain how outside reactions to the community have shaped its perspectives. The description of the community in this section provides context for the members' views on harassment and free speech in the next section.

User Interests

I used the Reddit data collected using PRAW to find the subreddits that were popular among KiA users. Table 3.2 shows the ten subreddits where the highest number of postings were made by these users. The popularity of subreddits like "MensRights," "The_Donald" (a subreddit for discussing Donald Trump[4]) and "politics" shows that many KiA users are interested in political and social issues that go beyond gaming culture. The subreddits that provide informative gaming content and discussions like "pcmasterrace" (a subreddit for PC gaming enthusiasts) and "Games" are also popular.

[4] Data was collected during the 2016 Presidential campaign in the US.

Table 3.2: Subreddits used by a sample of KiA submitters

Subreddit	**Number of Postings**
AskReddit	6072
MensRights	6047
pcmasterrace	5843
TumblrInAction	5126
The_Donald	4021
worldnews	3535
politics	3415
news	3223
GGFreeForAll	3064
Games	2943

I analyzed the flairs that KiA submissions were tagged with. Table 3.3 shows the results of this analysis. The predominance of flairs such as “Opinion” and “SocJus” reflects a focus on events and issues that are misreported or under-reported in the mainstream mass media or gaming media in the community’s view. Each time an event occurs that violates the principles of the community, the KiA platform works to report it, attract public condemnation of the event, and fuel subsequent action.

Table 3.3: Flair Distribution of a sample of KiA submitters

Flair	**Description**	**Count**
Opinion	Opinion pieces by mainstream media outlets or individuals, both positive or negative.	80
Humor	Jokes, memes, parody articles, etc.	74

Table 3.3: Flair Distribution of a sample of KiA submitters

Flair	Description	Count
SocJus	Relating to social justice affecting wider nerd culture.	67
Discussion	Serious discussion on a topic or question.	50
Ethics	Major findings on unethical press behavior.	42
Censorship	Censorship of GamerGate discussion or of other gaming-related issues.	31
Misc.	Anything else with a tangential relevance to GamerGate.	29
Industry	Relating to the wider games industry and issues within.	20
Drama	News of personal conflict between individuals, often GamerGate key players.	15
Meta	Relating to internal KotakuInAction affairs.	14
Meetups	Relating to offline GamerGate meetups.	12

I also analyzed the website domains of KiA submissions and found that 73.3 percent of submissions with links pointed to YouTube, Twitter and Imgur. Very few submissions pointed to news sites like New York Times and Washington Post. This shows the community's reliance on more informal channels of news.

Demographic Diversity

It is difficult to empirically determine the demographic diversity on KiA because of the pseudonymous nature of Reddit. However, through my observations, I identified some basic demographic indicators. I found that the ethos of the community is androcentric. Many users appear to have a libertarian attitude to society and culture. The threads on KiA often engage in discussions about American culture, media and politics, which indicates that a large number of users may be Americans.

Brad Glasgow conducted a survey of 725 individuals who support GamerGate on Reddit, Twitter or Voat (*GamerGate*). He found that 89.2 percent of the respondents were male. 25.4 percent were between 16-25 years old, 55.4 percent were 26-35 years and 19.2 percent were more than 35 years old. 74.5 percent of the respondents were white, and 8.8 percent were of Hispanic or Latino origin. 52.7 percent were Americans. Although I did not conduct or find a survey like this on KiA specifically, these results provide some insights into the demographics of KiA.

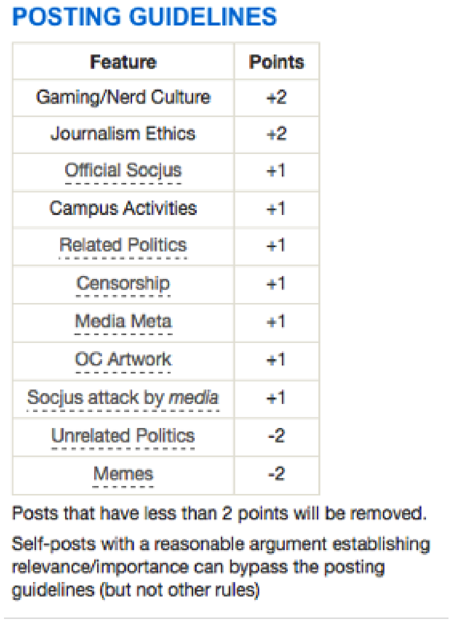
Goals and Beliefs

KiA provides a platform to its members where they can share and discuss information. It also provides a space where they can organize and mobilize supporters to take actions for addressing issues like media cronyism and censorship in games.

KiA users have diverse opinions. They often have different perspectives on how the community should operate. In mid-2015, the community split over the argument of what its focus should be. Some users believed that KiA should remain focused on its original goal of improving the standards of ethics in gaming media[5]. Many others argued that PC culture and third-wave feminism were responsible for many problems that pervade not just the gaming industry, but also the society more broadly. They insisted that the community should expand its objectives to fight against PC culture. In early 2015, the latter argument won, and the community expanded its focus. However, many members, including P12, a prominent KiA moderator at the time, left the community in protest.

In early 2017, after a community discussion, the moderators added a "points system" for new posts, to focus more strongly on core issues about gaming and nerd culture (see

[5] KiA's idea of "ethical media" constitutes games journalism that ensures disclosure of relationships between game developers and reviewers, if any, and guarantees that games are reviewed according to the principles of journalistic objectivity (Meditations, 2016). Some KiA users assume that any injection of politics in game reviews is unethical whereas others consider it appropriate if the subjective biases of the reviewer are disclosed.



POSTING GUIDELINES

Feature	Points
Gaming/Nerd Culture	+2
Journalism Ethics	+2
Official Socjus	+1
Campus Activities	+1
Related Politics	+1
Censorship	+1
Media Meta	+1
OC Artwork	+1
Socjus attack by *media*	+1
Unrelated Politics	-2
Memes	-2

Posts that have less than 2 points will be removed.

Self-posts with a reasonable argument establishing relevance/importance can bypass the posting guidelines (but not other rules)

Figure 3.1: KiA Posting guidelines

Figure 3.1) (ITSigno (submitter), 2017). A post must have two points or it will be removed by the moderators. “Related politics” is defined as relating to the Internet, gaming, and censorship. The exact details of the points system remain a topic of active debate, with changes made by votes of members.

Based on my interviews and observations, the community highly values freedom of expression, and opposes censorship in games and online communication. Many KiA members see themselves as average people and they consider GamerGate a grassroots movement to share their opinions and information about inept gaming companies and to bring about a change in the gaming industry and media.

The community also holds that the political correctness has gone too far in our society. KiA’s conception of political correctness can be described as what they consider a relentless push by SJWs to politicize every aspect of our society (especially gaming culture) and to accuse their opponents of holding views that could only be motivated by misogyny and bigotry. KiA users believe that they should actively work to fight against this PC culture. Participant P11 said that the community values encouraging an environment where people don't "feel like they need to walk on pins and needles to avoid offending others." Participant P3 said:

> *"What KiA stands for is that censorship in general, in art or if it's a person - that should not be censored just because you get offended and, of course, within the law. So, if it's an art that gets censored like games - that's something people usually are very against" (P3)*

Members believe it is their duty to oppose "social justice warriors" (SJWs). In the KiA world, an SJW is "*someone who has gone completely off the deep end in terms of ideology and logical fallacies. It's us versus them. They're always right no matter what*" (P13). Many KiA members accuse SJWs of being self serving, but insist that their own activism is selfless and legitimate. Participant P10 said that it is impossible to disagree with an SJW in any way on any topic, because, "*As soon as you disagree with them, you are a misogynist, you are a racist...you hate women, you are a rape apologist.*" Members of KiA believe that they are often falsely accused of a panoply of anti-social beliefs and actions. While the term "GamerGate" is synonymous with "hate group" to many outsiders, to KiA members, it is a positive affiliation to a group with values that challenge the status quo in a constructive way.

The KiA community tends to respond with anger to anything they see as attacking gamers. However, there is some disagreement about what constitutes legitimate criticism. One KiA poster writes about criticism of games, "*Some opinions are really against [all] video games and they're stupid. But some opinions are legitimate criticism of trends in game design. 'I'd like more non-violent games' is a legitimate opinion. But how to tell the difference*" (KotakuInAction, 2015)? What constitutes fair criticism (either of them or by them) is a controversial issue for the community.

Many participants stated that they use KiA to receive news on GamerGate and social justice issues. Some noted that they use KiA to organize "real-life" meetups with other KiA members. Many members believe that the media is unfair to the community, and KiA serves as an alternate news source by providing content that they do not find on traditional media. Other KiA users believe that the community serves as a "media watchdog."

> *"It's kind of, just sitting and waiting for crap to happen. Just to watch and call out, when media is acting up, when they are lying and, you know, pointing out the nepotism." (P10)*

Archiving

The community believes that record-keeping is important, and uses a number of tools to preserve records. One of the rules of KiA recommends its members to "archive as many things as possible." Archiving preserves articles in their original format, so that alteration or disappearance of embarrassing records from websites can be exposed, and media can be held accountable. 'Archive.is', a website that allows its users to save a text and a graphical copy for any webpage, is used for this purpose. One KiA user created "mnemosyne," a bot that automatically archives each submission on KiA.

The community also has an active blacklist of websites, and a bot automatically filters postings of submissions that link to any of these sites for review by the moderators before being posted. The submitter is also notified that this review can be bypassed, if the article is archived and resubmitted using the archived link. Members believe that the websites on this blacklist feature articles with sensationalist headlines related to GamerGate, so that they attract visits, and generate online advertising revenue. By using archiving, the community denies "click revenue" to these websites. Although KiA is firmly against censorship, the use of such practices on KiA indicates that members view their own moderation practices as quite different from those by others.

Each time the moderators take any action, a bot automatically posts the decision and a link to the KiA page where the decision occurred to a feed on modlog.github.io, an external website dedicated to building feeds of Reddit moderation. The deleted links are also tweeted on a public Twitter handle "@KIADeletedLinks." Though I have no detailed accounts of why this practice was adopted, the Twitter handle mentions “KiA’s Transparency Pledge.” I suspect that this practice represents the political value of transparency in governance.

Conventions and Policies

The activity on KiA is guided by (1) a set of established rules defined by Reddit guidelines; (2) a set of emergent rules that are specific to KiA; and (3) the norms of KiA. I discuss each of these in this subsection.

Conventions and Policies: Reddit’s Content Policy

Among other rules, Reddit’s content policy dictates that users are not allowed to post content that “threatens, harasses or bullies or encourages others to do so” (Reddit.com, 2016). The policy also describes how its rules are enforced, and ways of enforcing include banning of Reddit communities. KiA, like all subreddits, is governed by this policy. Some participants said that the biggest concern of KiA moderators is that the subreddit would get banned under this policy.

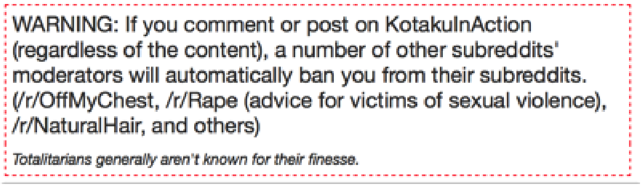
> *Outside of our little niche, there’s a lot of misinformation and general just dislike of us, and the admins are probably always looking for good reasons to ban us. (P6)*

The policy also states that individual subreddits may have their own additional rules. One instance where such local rules are enforced is when some subreddits like “Rape” and “BlackHair” ban any accounts that post submissions or comments on KiA (Figure 3.2). One member said that there are other subreddits (e.g., r/GamerGhazi) where a user gets banned if he claims his support for GamerGate. Some participants told me that they considered these policies unreasonable.

Conventions and Policies: KiA Rules

KiA has its own additional rules. The sidebar of KiA highlights these 9 rules[6]: “(1)

[6] These were the rules on KiA at the time of data collection. These rules continue to evolve over time.



WARNING: If you comment or post on KotakuInAction (regardless of the content), a number of other subreddits' moderators will automatically ban you from their subreddits. (/r/OffMyChest, /r/Rape (advice for victims of sexual violence), /r/NaturalHair, and others)

Totalitarians generally aren't known for their finesse.

Figure 3.2: KiA warns users when they post or comment

Don't be a dickwolf[7]; (2) No "Personal Information"; (3) No Politics; (4) Please tag posts for flair; (5) We are not your personal army; (6) Archive as much as you can; (7) Don't post bullshit; (8) No Reposts; (9) No MetaReddit Posts."

One KiA user interpreted the rule "Don't be a dickwolf" like this: "It means say whatever you want, but don't start hurling insults at fellow KiA members when you don't like what they say." The rule "We are not your personal army" forbids brigading[8], dogpiling (described under 'KiA and Online Harassment' section) and creating call-to-arms posts against individuals. Links to comments of other subreddits are automatically banned by a bot. The "Don't post bullshit" rule prohibits users from posting "editorialized headlines and links to provably false information" (Kotaku In Action, 2016). It urges members to provide information without trying to spin a narrative.

Conventions and Policies: KiA Norms

The norms of the community dictate that moderation should be minimal. Many members strongly believe in freedom of expression. There is a spectrum of interpretations of limits on free speech, and KiA leans towards favoring no limits. Many participants expressed their concerns about social media platforms shutting down parts of the political spectrum by censoring selected conversations or banning certain users. The community claims that it values discussion, and it believes that everyone should be allowed to have an equal voice. Members consider that this belief guides the norms of the community, where users are not banned if they express an unpopular opinion.

> *"I think the correct term would be "laissez-fair," the kind of hands-off mod-*

[7] The word "Dickwolf" originated in a controversy over a 2010 comic strip (Fudge, 2013).

[8] Brigading is a concerted attack by one online group on another group, often using mass-commenting.

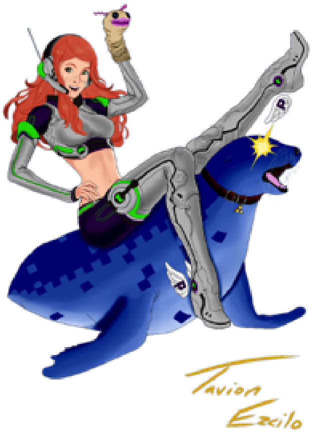


Figure 3.3: KiA Header Image

> *eration that allows us to really post anything that is related to our pretty lofty generalized goal. So you can get any movement going as long as it's kind of related, is one big benefit." (P6)*

However, the voting mechanism of Reddit does not allow posts with unpopular viewpoints to appear at the top. Some participants admitted that they do not often see anti-GamerGate posters on KiA, and even when such users show up, their posts frequently get down-voted, and thereby buried under the more popular pro-GamerGate posts.

The moderators try to find a good balance between freedom of speech and on-topic discussions. There are ways to get banned or to have a post deleted, but such penalties are not likely the result of using unacceptable terminology or expressing an unpopular point of view. Even though the moderation is minimal, it still has a significant impact. For example, Participant P13 posts content on Twitter that he says he is sure would never be allowed on KiA.

The community embraces a few norms and practices that are widely condemned by outsiders, and some of these norms have emerged in response to the reaction of outsiders to the community. Consider that the primary visual draw on the KiA subreddit header is the image (Figure 3.3) of a red-haired woman named Vivian James, a frequent character in GamerGate-related comics. This image depicts Vivian riding a sea lion while wearing a sock-puppet on one hand. This reference two separate practices that are important to the community:

(1) “Sock-puppeting”: This refers to the creation of auxiliary accounts by a user to provide anonymous support to arguments posted by his main account. Anti-KiA communities often blame KiA for engaging in identity deception using sock-puppeting. KiA strongly refutes such allegations, and uses a sock-puppet in its header image to poke fun at them.

(2) “Sealioning”: This refers to the act of persistently but politely requesting evidence in a conversation. The name “sealioning” comes from a comic in which a sea lion annoys a couple by trying to engage them in a conversation that they are not interested in having (Malki, 2014). Sealioning is viewed by anti-GamerGate users as intrusive attempts at engaging an unwilling debate opponent and excessively requesting evidence. However, KiA has cultivated and embraced “sealioning” as a rhetorical norm in which members practice providing and requiring evidence and source material while engaging in regular conversations. As I will discuss later, the presence of trolling might have encouraged this practice too. A notable example of collecting evidence is the frequent use of the site deepfreeze.it, on which KiA users and other GamerGate supporters document evidence of unethical gaming journalist behavior for use in later arguments.

3.4.2 KiA and Online Harassment

I began this research with a set of questions about online harassment. Do KiA members engage in harassment? What do they think “harassment” is? In this section, I discuss participants’ response to accusations of harassment.

I asked many of my interview subjects what they thought “harassment” is. I also analyzed discussions about online harassment on KiA. A consensus for the definition and resilient characteristics of harassment from the community’s perspective emerged from the collated responses. Participants divided communicative acts by their intensity. Intensity has two components: the content of an individual message, and the frequency with which messages are sent.

Content

Participants felt that a single message may or may not be harassment, depending on the content:

> *"I think that a big problem is the redefinition of what harassment means. I think. . . just sending mean tweets is considered harassment, as opposed to just like telling someone to explicitly kill yourself. . . there's definitely different degrees." (P2)*
>
> *"I think, GamerGate sees harassment in the same way that any average, ordinary person or even the law views harassment - which is you know, stalking, death threats, as in calling your house; a very persistent, when people are persistent in their stalking, or harassment, basically the way the law sees it." (P10)*

This indicates that in some instances, KiA members may not grasp the emotional toll that their remarks can exert on other users. As Whitney Phillips describes, "even the most ephemeral antagonistic behaviors can be devastating to the target, and can linger in a person's mind long after the computer is powered down" (Phillips, 2015b).

Many participants expressed their concerns about conflation of criticism with harassment. Some claimed that opponents of GamerGate use accusations of harassment to fight against differences of opinions. They felt that a critical consequence of this conflation is affective desensitization of many users in the community to the concept of harassment.

> *"It is ultimately derailing 'cause the entire thing has been, we're trying to have one conversation and the other people just call us sexists and harassers, and it's like, 'that's not a response.' " (P2)*

> *“The accusations, the racist and misogynist, that is their one go-to insult to shut down anything that you have to say.” (P10)*

Some participants said that KiA opponents sometimes consider expressions of sincere disagreements with them as harassment and block them.

> *“You could be in a debate with somebody else, and you could be asking for what are the gun statistics from 1980, and they’ll be like "That’s a sealioning question. I don’t trust you. You’re blocked” . . . and I’d be like, what the fuck just happened here?” (P13)*

Frequency

Some participants felt that postings by a sole user on a single social network should not be considered harassment, since almost every social network provides its users the ability to ‘block’ accounts. When a user blocks an account, the blocked account can no longer send messages to the user. Participants argued that harassment entails a more persistent behavior, where the harasser is willing to create new accounts when banned or blocked, and continues to send threats to the victim.

> *“Harassment requires a little more motivation, a little more intent, a little more longevity. Someone kind of drunkenly wandering up to your house in the middle of the night and knocking on your window isn’t stalking. But doing it 20 times might be.” (P6)*

Another scenario is that of “dogpiling.” KiA rules explicitly prohibit calls for dogpiling, which is an indication that the community has had previous trouble with dogpiling. Dogpiling occurs when a single individual is overwhelmed by receiving a large number of messages from different people. Such messages are often sent through private channels on social networks, and therefore it is not always obvious to the sender that the target is receiving hundreds of similar messages. Although the content of individual messages may not be seriously threatening, the receiver feels vulnerable and threatened.

Justification

Many members also distinguished among different communicative acts by whether they considered the act to be an appropriate response to a prior offense. Some participants observed that individuals who commit social transgressions in public deserve to be called out for their actions.

> *"You're in a public space. If you're acting like a child, if somebody films you, that's kind of your fault." (P13)*

A few felt that such actions also include the postings made on websites like Twitter. They argued that such websites should be considered public spaces because the content on them is publicly available.

Many members also took into consideration the status of the account receiving attacks. They felt that individuals who are "public figures" should expect to lose certain protections. Participant P13 distinguished between messages sent to identifiable, personal accounts versus those sent to anonymous or organization accounts.

A few participants discredited harassment claims, and said that harassment victims were soliciting their own abuse. Participant P2 noted that some people provoke angry responses by saying something inflammatory, and then present responses as evidence of harassment.

> *"My personal opinion is that it seems like there is a weird culture of victimhood, where victimhood is put up on some kind of weird pedestal." (P4)*

Accountability

Some participants felt that in a large, leaderless community like KiA, it is difficult to rein in every user, and problems like harassment are bound to occur. They assume that the nature of the Internet and online social networks makes harassment in any contentious online movement inevitable.

> *"I think this is the biggest thing the media fails to get about Gamergate: it's basically the same as the rest of the Internet outside Gamergate. The Internet is an incredibly hostile place sometimes." (P9)*

A few participants admitted that some users who subscribe to KiA have engaged in harassment. However, they claimed that such users do not represent the values of KiA and are at the periphery of the community. They challenged the accusation that organized harassment exists in the community.

> *"Do I think that there are still elements of KiA who support harassment? Yeah, of course. I actually had an argument the other day with someone who wanted to get somebody fired." (P5)*

Perceptions of Unfair Portrayal

Many participants felt that it was unfair of the media and members of other online communities to label everyone associated with KiA as harassers. They believe that an overwhelming majority of KiA members act responsibly and do not cause any problems. In their view, most of the users in the community strongly condemn harassment and doxing, and many work hard to ensure that such activities do not occur from inside the community. One former KiA moderator said:

> *"Reading what was being said about KiA, that it was a front for an abuse campaign, was enraging. The mod team had done all we could to keep that sort of stuff out of the sub, and to discourage it at every possible avenue." (P12)*

Many participants said that a number of users in KiA also got harassed and received death and rape threats, but such incidents were dismissed by the media and KiA opponents.

In their view, the media deliberately under-reported such incidents to spin the narrative of KiA as a hate group, and undermine its arguments about ethics in game journalism.

> *"Let's not forget that there is a lot of talk about harassment of these female developers and these prominent people, but there are a lot of people who supported KiA who were harassed, who were doxed and who were kind of blacklisted because of their support for this, basically this idea." (P5)*

Doxing

KiA's 'No "Personal Information" rule' bars users from sharing individuals' phone numbers, addresses, and other private information, and asks them to avoid posting links into people's Facebook pages. A few participants stated that they haven't seen any incidents of doxing on KiA. Others mentioned that they have rarely seen doxing, but ignored such posts.

> *"Most people don't care. And so somebody says like this is a person I hate and this is his home address, you don't care. You would look away. You would be slightly disgusted. That's how most people felt, so they didn't look at it." (P4)*

Participant P8 claimed that doxing is likely to occur only to users who are known over several platforms, and are "something of a public figure." It was common for many participants to state that the claims of harassment and doxing were largely exaggerated.

Participant P4 investigated some incidents of doxing, and found that they were being organized by a troll group. Some participants pointed out incidents where KiA users got doxed. Participant P12 said that his personal information was revealed, but the source of doxing information was promptly suspended, and the information was quickly removed.

Some participants claimed that they have seen information presented as evidence of doxing on KiA that they wouldn't consider doxing. This raises questions about the boundaries of what should be considered doxing. For instance, if information is readily available using an Internet search engine, should it be considered doxing? One moderator explained how KiA deals with incidents of doxing:

> *"We have a pretty limited set of tools for when that happens. Naturally - if people post dox on KiA, we remove it. If someone decides to post personal information on another site, there's very little we can do." (P8)*

A few participants pointed out the existence of a "harassment patrol" in the community's early days. It consisted of a group of users who actively looked out for trolls on the community, and took actions to ban them. They also reported doxing incidents and prevented users from organizing brigading activity on the community.

Trolling

KiA hosts a variety of fast-moving discourse that includes good-natured ironic posts, humorous or sarcastic comments, pranking and sensationalist exaggeration. The more antisocial aspect of this discourse is trolling. Trolling entails provoking others to engage in pointless, time-consuming discussions (Donath, 1999; Herring et al., 2002; Kraut and Resnick, 2012).

In her study on Internet trolls (Phillips, 2015b), Whitney Phillips notes, "Trolls believe that nothing should be taken seriously, and therefore regard public displays of sentimentality, political conviction, and/ or ideological rigidity as a call to trolling arms." The strong ideological stances on both sides of the GamerGate controversy make it an attractive target for trolls. They often disrupt the discussion space on KiA and other GamerGate-related forums. Some participants also accused such users of "false flagging" (falsely blaming KiA users for operations that they did not conduct).

A few participants talked about troll groups, some of which are splinter groups that emerged out of the GamerGate forums on 4chan and 8chan websites, that attempt to disrupt and mislead discussions about GamerGate. Some “third-party trolls” harass members of both GamerGate and anti-GamerGate communities, and blame the other group, to instigate the groups to fight each other. GamerGate supporters have often claimed that “much of the mayhem associated with the movement comes from third-party trolls who get a kick out of baiting both sides” (Young, 2015).

A few KiA users have called for separation from #GamerGate and rebranding under a new tag to distance the movement from its association with harassment in the popular zeitgeist. Such calls were overwhelmingly rejected by the community because it felt that any rebranding would divide the community, and the trolls would simply follow the movement and smear it under the new tag.

The presence of these trolls might have affected the discourse on KiA. For instance, users are often asked to provide evidence to back up their claims, so as to ensure that they are not trolling. The KiA rules also prohibit posts and comments that “are clearly not intended to generate discussion, but rather just aimed at generating as much drama and outrage as possible” (Kotaku In Action, 2016).

Legal Discourse

Some participants mentioned that instances of harassment and legitimate harm should be taken seriously, and such instances should be reported to and handled by authorities.

> *“If it’s actual harassment, go to the police. That should be the end of it. But they make the argument that police don’t do enough to help, and that’s probably true. But, they probably have more pressing things than somebody bothering someone on the Internet.” (P2)*

Participant P2 expressed concern that illegitimate cases of online harassment might discourage the authorities to deal with legitimate cases in the future.

All of my participants said that they do not personally engage in harassing others, and that KiA as a group specifically prohibits all forms of harassment. This is in sharp contrast to the portrayal of GamerGate in the popular press. There are a number of possible explanations for this apparent contradiction. First, there is a self-selection bias in how my interview subjects were selected. The people who would agree to speak to academics about KiA are likely not the ones doing the harassment. Second, because KiA is a large, leaderless group, it is impossible to hold the group accountable for the behavior of any single individual, because that individual is simply redefined as not speaking for the group (like the "no true Scotsman" logical fallacy (Fieser and Dowden, 1995)). Third, underlying this contradiction is a sincere disagreement about what is "harassment" and what is free speech. A fourth potential explanation is that my informants were simply lying to me. Although they are clearly putting their best selves forward, I believe them to on the whole be giving honest accounts. The other three explanations are all true in varying degrees. I conclude that KiA is not a viper's nest, though there are probably vipers in the nest.

3.5 Discussion

As we have seen, KiA members believe that they have been wrongly accused of harassment. The other side of the controversy, of course, has a radically different account of what has actually taken place. It is indisputable that both sides have both experienced and committed harassment. What is impossible empirically for me to determine is the relative prevalence of harassers in the community. My informants state that a few bad apples are giving the entire group a bad name. Others label KiA a hate group and insist that it is immaterial that there are a small number of decent people mixed in. The relative prevalence is an empirical question that I cannot answer, and nothing in this study should be construed to support one view or the other.

In her book "Hate crimes: Causes, controls and controversies," Phyllis Gerstenfeld asserts that there is no simple way to define a hate group, and "whether a particular group is to be classified as a hate group is often in the eyes of the beholder" (Gerstenfeld, 2013). She argues that one of the problems with identifying hate groups is that "some organizations have certain factions that are clearly bigoted although other factions are not."

When a movement is made up of people with differing views and tactics spread across multiple websites, it is impossible to hold the group accountable for the action of any individual, because the individual's actions can always be redefined as not representing the group. Trying to hold an entire loosely defined group responsible for the actions of its worst behaved members appears to be a catalyst for escalating rancor on all sides. Therefore, we should encourage efforts to understand the commonly held values of such groups instead of characterizing them by the views and actions of outliers.

3.5.1 Implications for Theory: Free Speech vs. Harassment

Prior research has suggested that anonymity provided by Reddit, Twitter and other social media websites lowers social inhibition, and encourages users to be more aggressive in their communications (Kraut and Resnick, 2012). This leads to situations in which some users see their online behavior as innocuous or an exercise in free speech, but it is construed as online harassment by others.

In her study on cyber-racism, Jessie Daniels notes that there is a US/Europe cyberhate divide (Daniels, 2009). She explains that the US response to white supremacy online is to view it as speech protected under the First Amendment and to forfeit it only when it is joined with conduct that threatens, harasses, harms, or incites illegality. In contrast, other Western industrialized democracies address online racism by broadening the scope of their existing antiracism laws.

There exist similar disagreements on the questions about the content of other hateful material. The enormous international influence of the US policies and its prominence as a safe haven for hosting Internet hate speech reduce the likelihood that nations who wish to regulate hate speech online will be able to do so. Besides, as Titley et. al point out, "there seems to be consensus that the problem of cyberhate is increasing both in magnitude, and in the variety of strategies used" (Titley, Keen, and Földi, 2014). This reflects a need to consider analytic alternatives to the binary interpretation of free speech versus harassment that many KiA users seem to hold. Distinguishing controversial speech from hate speech, and weighing "freedom of speech" against protection from abuse would help the researchers and regulators think more critically about these issues.

These tradeoffs are further complicated by the presence of trolls who pretend to be sincere members of the community and lure others into pointless discussions. Therefore, efforts to characterize the evolution of troll behavior along the lines of Whitney Phillips' work on trolls (Phillips, 2015b) should be encouraged so that platforms can efficiently identify trolls and regulate their postings.

Some researchers argue that the internet has witnessed a number of moral panics regarding online activity, and this clouds the fact that only a small minority of users actually engage in disruptive or illegal activity (Ellison and Akdeniz, 1998). My findings indicate that there may be an aspect of moral panic in response to GamerGate. Many KiA users believe that they experienced one-sided reporting by the media. They argue that misinformed reactions and stereotyping by their opponents fueled the anger of GamerGate supporters and intensified their activities. To break this cycle of negative reinforcement and escalation, our findings suggest that it is advisable to take a more balanced tone in response. If outsiders seek to find common ground with the more moderate members of the movement, this can break the cycle of escalation.

While outsiders might seek mutual understanding with the more moderate members of the group, it is clearly not ethical or strategic to appease the true harassers in any way. But how do we tell the difference?

My findings suggest two key dimensions to understand controversial speech: its intensity, and whether it is perceived as justified from a particular perspective (see Table 3.4). Intensity has two dimensions: the strength of each utterance, and how often the communication is repeated.

Table 3.4: Conceptualizing Controversial Speech

	JUSTIFIED	UNJUSTIFIED
HIGH INTENSITY	Public Shaming	Harassment
LOW INTENSITY	Criticism	Insult

A single utterance can qualify as "high intensity" if it is strongly worded or contains an actual threat. On the other hand, a simple utterance (like "you're wrong") might be perceived as intense if it is repeated many times. As I discussed in my findings, some users may feel harassed when they receive a large number of messages, even if the content of individual messages may not be seriously threatening from the sender's perspective. In such situations, it may not be reasonable to label all message senders as harassers, even though harassment has occurred.

When people feel that their intense criticism is justified, the activity is often called "public shaming" rather than harassment. As Jon Ronson has thoughtfully documented (Ronson, 2015), public shaming often has consequences for the individual (like loss of job) that outweigh the perceived offense.

One key point to note here is that both intensity and justification are subjective. Underlying much of the controversy I observed are *disagreements about what quadrant we are in*. One person's "criticism" is another person's "harassment."

Judgments of intensity differ radically depending on an individual's basic views on the proper limits on free speech. Citron (2014) writes that although online speech is crucial for self-government and cultural engagement, certain categories of low-value speech, e.g., true threats, defamation, fraud and obscenity, "can be regulated due to their propensity to bring about serious harms and slight contribution to free speech values." Everyone agrees that you can't yell "fire" in a crowded theater. Beyond concrete and immediate harm, where do we draw the line? Feminists and critical race theorists argue that words have power, and we are responsible for the emotional harm our words may cause others (Daniels, 2009; Spender, 1985). Strong civil libertarians argue that censorship is a slippery slope, and freedom of speech includes the right to offend (Brennan, 2012).

With such fundamental disagreements on what sort of speech is appropriate, it is a wonder that people ever succeed in civil communication. Fortunately, this problem is normally solved by social norms. Members of different online communities develop a sense of local social norms for appropriate communication. Each subreddit in fact evolves its own norms—what you may say on Kotaku in Action is quite different than what you can say on GamerGhazi or AskReddit or any of the other thousands of subreddits. Conflict about appropriate speech is particularly likely to emerge on sites like Twitter, where it is not clear whose social norms apply.

3.5.2 Implications for Design

A key approach to managing the problem of online harassment is by developing moderation and blocking mechanisms (Crawford and Gillespie, 2016; Geiger, 2016; Lampe and Resnick, 2004). My findings add nuance to our understanding of the challenges of this undertaking. As I discussed, the tradeoffs between online harassment and free speech are complex. Couching too broad a spectrum of online dispute under a single umbrella of harassment can lead to broad reactionary interventions that are problematic. Although it may appear that laying out detailed, formal rules to guide moderation would bring a sense of fairness in an online community, the moderators need enough flexibility to judge any action in its context. I will reflect on this need in my description of the use of automated tools for content moderation in Chapter 5 again. Moderation decisions should take into account the intensity of the language used, as well as the frequency of communications directed at a single target. As Phillips recommends, moderation decisions should also consider the persistence and relative searchability of data for a given behavior (Phillips, 2015b). Supporting a personalized approach to controlling the user's social feed should also be encouraged.

I argue that another, complementary direction where designers can focus is the design of tools that can help improve discussions and mutual understanding of groups with different ideologies. This is a challenging problem. For example, consider the context of GamerGate. It is difficult to create a legitimate dialogue between the two sides: Basic language choices (for example, KiA's use of the phrase "social justice warriors") posit deep-seated assumptions about the other side. The opponents of GamerGate view it as a hate group, while its supporters believe that their legitimate concerns are rebuffed by portraying them as harassers. I propose that one useful way to address such challenges is to draw from prior research on modeling argumentation for the social semantic web. There is a vast body of work in this area where researchers have proposed theoretical models and implemented social web tools that help users engage in argumentative discussions (Schneider, Groza, and Passant, 2013). For example, Kriplean et. al. developed ConsiderIt, a platform that allows users to author pro and con points (Kriplean et al., 2012). This augmented personal deliberation helps mitigate the opportunities for conflict that occur in direct discussions while allowing users to consider the arguments on the other side.

Differences in cultural norms among groups can make communications difficult. It would be helpful to design solutions that help bridge across different norms of politeness. Disentangling the mode of address from content can help. Grevet's work on designing social media to facilitate more civil conversations provides useful insights (Grevet, 2016). Such tools can help users identify common ground.

3.5.3 Limitations

In this research, I deliberately sought out just one side of the controversy: Who are the people on KiA, and how do they view their activities? I am not attempting to make any statement in favor of or against members of KiA, but simply to try to see what the world looks like from their point of view. What I found had much more complexity and nuance than I originally anticipated. In the future, it would be interesting to study individuals who oppose GamerGate on Twitter and sites like GamerGhazi.

A key limitation on my findings is the nature of my sample. The sample size is small but I triangulated the interview data with notes made during participant observation on KiA. I note a self-selection bias — my research team only spoke with KiA members who were willing to talk to us. Additionally, social desirability bias might have motivated our interviewees to under-report behaviors that may be viewed as unfavorable.

I hope that this study has been fair even though, given the size and diversity of the KiA community, there may well be important exceptions to what I have described and observed.

3.6 Conclusion

What made Lindy West's story so compelling is that she and her harasser transcended their differences and reached a degree of mutual understanding. Nothing like that happened in this study. I have no reason to believe that anyone we spoke to developed any new insights into how their actions might affect others. I may perhaps have helped outsiders to develop some understanding of our subjects and their concerns. Members of KiA have a mix of concerns some of which a neutral outside observer might find reasonable to a degree (like frustration with political correctness, frustration with policies of the game industry, and concerns about the quality of journalism), and other concerns likely less so. However, dismissing their concerns entirely simply fuels their righteous anger. The path to greater harmony is through mutual understanding. As it was for West and her harasser, that understanding needs to be two-way. The intriguing question for the research community is whether it is possible to design tools and systems to help foster such understanding.

In this work, my research team and I have spoken with a self-selected subset of members of the KiA subreddit who were willing to speak with academics, and were likely on their best behavior while doing so. However, even from observing this tiny slice of the broader community, I found fascinating and unexpected complexity and nuance.

Clifford Geertz writes that anthropologists don't study villages—they study *in* villages (Geertz, 1973). Studying in this particular techno-village, I observed, first, a fundamental problem of accountability in distributed social movements. New tools that help visualize the beliefs of groups might help outsiders distinguish common beliefs from rare ones. Second, I find that what is "harassment" is often in dispute. Our informants complain that simple disagreement on their part is often portrayed as harassment. It is difficult to resolve issues of possible harassment if we cannot even agree on whether it is taking place. Different design solutions may be needed for addressing deliberate harassment versus sincere misunderstanding and communicating across different social norms of conversation. Although much work has been done on blocking tools for deliberate harassment, there are a host of open research questions about how to create tools to support greater understanding. My findings suggest that barriers of language use and differences in social norms of politeness often obscure underlying common values, and these challenges may be amenable to designed solutions.

CHAPTER 4

UNDERSTANDING THE USE OF THIRD-PARTY MODERATION TOOLS: THE CASE OF TWITTER BLOCKLISTS

4.1 Introduction

4.1.1 Online Harassment

In mid 2016, 25-year-old Erin Schrode was in the middle of her congressional campaign. She was aiming to become the youngest woman ever elected to the U.S. House of Representatives. Days before the election, she began receiving hate-filled emails and tweets from anonymous individuals who targeted her for being Jewish. One email said, "Get to Israel to where you belong. That or the oven. Take your pick" (Associated Press, 2017). Another said, "... all would laugh with glee as they gang raped her and then bashed her bagel-eating brains in" (Pine, 2016). On election day, Schrode switched on her computer and found that her campaign website had been hacked and all references to her name were converted to "Adolf Hitler." Over the next few months, the attacks grew more numerous and repulsive. Some posters attached doctored photos of Schrode in their messages - one sent a photo of her wearing a Nazi style yellow star; another sent an image of her face stretched onto a lampshade. Every time Schrode looked at any of her social media feeds or emails, she was reminded that she was unwelcome and told that she was inferior. She often felt lonely and suffocated. "You read about these things in the news," she said, "but it's so unreal when it targets you" (Associated Press, 2017).

Schrode's experience is far from unique. In recent years, online harassment has emerged as a growing and significant social problem. According to a 2017 Pew Research study, 41% of American adults have experienced online harassment and 66% of adults have witnessed at least one harassing behavior online (Duggan, 2017). This study also found that social media is the most common venue in which online harassment takes place (Duggan, 2017). Many online offenders have turned social media platforms into forums to bully and exploit other users, threaten to harm or kill them, or reveal sensitive information about them online.

The problem of online harassment is particularly prevalent on Twitter[1] (Matias et al., 2015; Geiger, 2016). Some critics have worried that Twitter has become a primary destination for many trolls, racists, misogynists, neo-Nazis and hate groups (Warzel, 2016). Twitter has indeed found itself ill-equipped to handle the problem of online harassment, and its CEO has declared, "We suck at dealing with abuse and trolls on the platform and we've sucked at it for years" (Tiku and Newton, 2015).

In this chapter, I use Twitter blocklists, a third party blocking mechanism aimed at addressing online abuse on Twitter, as a vehicle to explore the problem of online harassment. I review different experiences and perceptions of online harassment. My findings show that many Twitter users feel that existing moderation tools on Twitter fail to provide them with adequate protection from online abuse, and they circumvent the limitations of such tools by developing, deploying, and promoting third-party moderation tools like blocklists. I investigate how the use of blocklists is perceived by those who use them and by those who are blocked because of them[2].

4.1.2 Blocking on Twitter

In this section, I explain the use of blocking and muting mechanisms on Twitter. Following this, I will describe Twitter blocklists.

Many platforms have implemented moderation mechanisms to discourage antisocial behavior such as trolling and harassment. These mechanisms include using a small number

[1] https://twitter.com. Founded in 2006, Twitter is a microblogging platform that allows its users to post 140-character messages, called tweets, about any topic, and follow other users to read their tweets. Recent news articles suggest that Twitter is one of the five most popular social network sites worldwide (Moreau, 2016). As of the first quarter of 2017, it averaged at 328 million monthly active users (Statista, 2017).

[2] Findings from this study were published in 2018 in the ACM Transactions on Computer-Human Interaction (TOCHI) journal (Jhaver et al., 2018). Sucheta Ghoshal assisted me and my PhD advisors, Amy Bruckman and Eric Gilbert, guided me on this work.

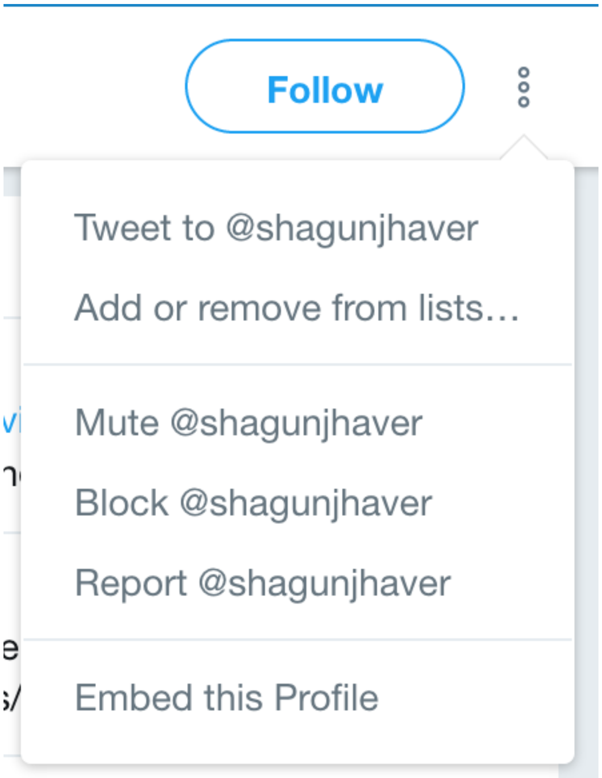


Figure 4.1: Twitter provides options to mute, block, and report offensive users.

of human moderators who manually remove abusive posts (Kiesler, Kraut, and Resnick, 2012), moderating through a voting mechanism where registered users up-vote or down-vote each submission (Lampe and Resnick, 2004), and allowing users to flag abusive content (Crawford and Gillespie, 2016). Another mechanism that many platforms primarily rely on is providing users the ability to mute, block, or report offensive users (Figure 4.1).

On most platforms, and particularly on Twitter, blocking or muting an account allows a user to stop receiving notifications from that account, and that account's posts don't appear on the user's timeline or newsfeed (Twitter, 2016a). The difference between blocking and muting is as follows: blocking an account prevents that account from viewing the blocker's posts or sending direct messages to the blocker. In contrast, a muted account can still view the user's posts, "favorite" them, and reply to them. Muting an account is more socially delicate than blocking it: a muted user is not notified that he is muted, and he may continue posting to the user who muted him without realizing the receiver cannot see his posts, whereas a blocked user immediately realizes that he is blocked if he attempts to post to the blocker. If the blocked user accesses the blocker's profile, he sees the following message:

"You are blocked from following @[blocker] and viewing @[blocker]'s Tweets. Learn more"

Blocklists are third-party Twitter applications that extend the basic functionality of individual blocking by allowing users to quickly block all accounts on a community-curated or algorithmically generated list of block-worthy accounts (Geiger, 2016). Perhaps the most popular type of blocklists are anti-harassment blocklists that aim to block online harassers en masse. The use of decentralized moderation mechanisms like anti-harassment blocklists takes some pressure off the centralized Twitter moderators so that they don't have to be as strict in their moderation. Everyone has slightly different boundaries, and the use of these lists can provide users an experience that is more customized to their needs (Geiger, 2016). However, not everyone who is put on anti-harassment blocklists sees himself as a harasser. Some of the users blocked by these lists may think of themselves as perfectly reasonable individuals. I will expand on this problem and other limitations of blocklists in my findings.

Next, I discuss two different Twitter applications that serve as blocklists.

Block Bot

Block Bot[3] was the first blocklist implemented on Twitter. It is a socially curated blocklist where a small group of moderators coordinate with one another and make complex decisions about which Twitter users to put on a shared list of blocked accounts.

Block Bot emerged out of the use of hashtag #BlockSaturday[4] on Twitter. In 2012, a Twitter user began identifying accounts that he felt were worthy of blocking and started

[3] `http://www.theblockbot.com`

[4] #BlockSaturday is a wordplay on a popular trend of using hashtag #FollowFriday. Twitter users used #FollowFriday in their posts to recommend to their friends on Twitter which other handles they should follow.

posting tweets containing the usernames of these accounts along with the #BlockSaturday hashtag. This was done so that his followers and anyone following the hashtag #BlockSaturday could block those accounts (BlockBot, 2016). As more users began posting such tweets and this trend became more popular, a few users expressed a need to automate the process of blocking. This led to the creation of Block Bot. Its developers made creative use of Twitter APIs that were developed to support third-party clients such as smartphone applications (Geiger, 2016). The use of this blocklist allowed users to collectively curate lists of Twitter accounts that they identified as harassers and block them together quickly and easily.

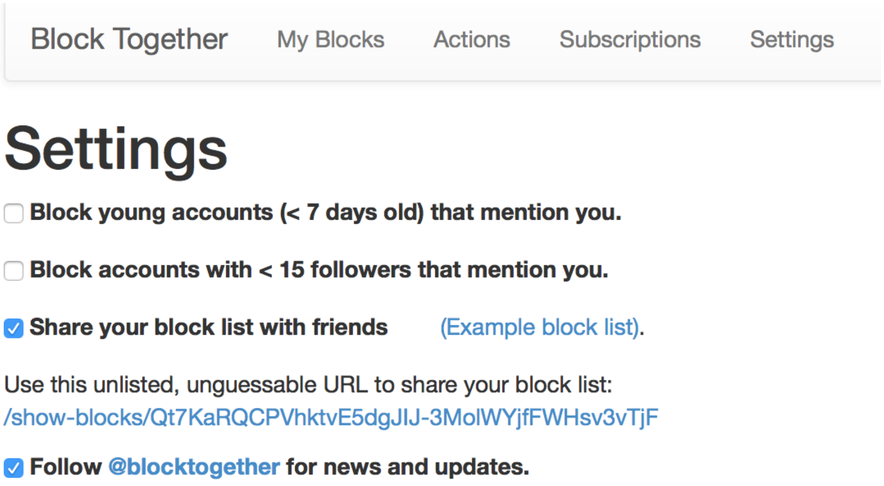
In its initial days, Block Bot was primarily used to serve the atheist feminist community and block individuals who opposed the rise of the Atheism+[5] movement. Block Bot later expanded its goals to block supporters of GamerGate movement (GamerGate Wiki, 2016), users who harass transgender people, and other abusive accounts. This blocklist allows moderators to sort blocked users into three categories of offensiveness – nasty, unpleasant, and annoying. It also allows subscribers to pick the level of offensiveness they would like to excise from their Twitter feeds (Hess, 2014).

Block Together

Block Together[6] is a web application that serves as a "centralized clearinghouse" for many blocklist curators and subscribers (Geiger, 2016). Like Block Bot, Block Together is a Twitter application that was developed by volunteers to combat harassment on Twitter. It was released by third-party software developers at the Electronic Frontier Foundation (Duggan, 2014). In contrast to Block Bot, which hosts a unique list of blocked accounts, this application hosts many different lists of blocked accounts. Block Together allows Twitter users to share their own list of blocked accounts that other users can subscribe

[5] Atheism+ is a movement that originated in August 2012 by blogger Jen McCreight. It encouraged progressive atheists to cater to issues other than religion, such as social justice, feminism, racism and homophobia (RationalWiki, 2016).

[6] https://blocktogether.org



Settings

☐ Block young accounts (< 7 days old) that mention you.

☐ Block accounts with < 15 followers that mention you.

☑ Share your block list with friends (Example block list).

Use this unlisted, unguessable URL to share your block list:
/show-blocks/Qt7KaRQCPVhktvE5dgJlJ-3MolWYjfFWHsv3vTjF

☑ Follow @blocktogether for news and updates.

Figure 4.2: BlockTogether settings for blocking accounts on Twitter.

to (Figure 4.2). It also gives the subscribers an option to block accounts that are newly created or have fewer than 15 followers. This helps combat trolls who create new accounts on Twitter after they find themselves being blocked. Although Block Together was created to address online abuse, it now hosts many blocklists that serve vastly different purposes, including those that block spam accounts and those that block ISIS critics. However, in this chapter, I restrict my discussion to anti-abuse blocklists.

Next, I discuss Good Game Auto Blocker, a popular blocklist that is hosted by Block Together.

GamerGate and Good Game Auto Blocker

Online harassment often occurs as a result of coordinated harassment campaigns organized by hate groups that overwhelm a target by synchronously flooding his or her social media feeds (Geiger, 2016). One group that has recently gained attention in the popular media for coordinating such campaigns is GamerGate[7]. Although conversations about

[7] GamerGate is an online social movement that emerged in response to a series of controversial events surrounding game developer Zoe Quinn (Jason, 2015). The supporters of the movement insist that GamerGate stands for ethics in gaming journalism. However, a number of media articles portrayed the movement as a hate group, and claimed that users supporting the movement engage in death threats, rape threats and doxing among other harassment activities. Mortensen provides a detailed account of the progression of events that helped GamerGate gain popular attention (Mortensen, 2016) .

GamerGate can be found on many online sites like Reddit (Chapter 3), Voat, 8chan and YouTube, Twitter has emerged as one of the most popular sites for discussing this movement. Many Twitter users who oppose GamerGate consider its supporters as harassers and feel a need to block them en masse. This led to the creation of Good Game Auto Blocker (GGAB), a blocklist aimed at blocking GamerGate supporters on Twitter.

GGAB uses Block Together to implement blocking on Twitter. The current procedure for blocking users on GGAB is unknown, but at least in its initial days, GGAB used a predominantly algorithmic approach to curate the list of block worthy accounts (GamerGate Wiki, 2017). It collected the followers for a short list of prominent #GamerGate contributors on Twitter. If anyone was found to be following more than two of these supporters, they were added to the list and blocked. This blocklist also used a periodically updated white-list of users who satisfied this criterion but were false positives. The accounts on the white-list were unblocked. For this blocklist, a block removal can be appealed by providing a link to one's Twitter profile and an explanation of why an exception should be made.

In this study, I interviewed users who subscribed to GGAB or Block Bot blocklist as well as those who are blocked on such lists in order to understand the motivations, benefits and limitations of the use of this novel socio-technical, anti-abuse mechanism. My findings indicate that many users find blocklists to be quite effective in addressing online harassment. However, the widespread use of blocklists, as they are currently implemented, can also lead to many problems. For example, I discovered the problem of a "blocking contagion": when a popular blocklist is forked to create multiple other lists, false positive accounts on the original blocklist end up getting blocked by users who subscribe to any of the several forked lists. This results in several users inadvertently censoring these false positive accounts. I discuss this phenomenon and other shortcomings of blocklists in my findings. I also discuss the advantages and limitations of using GGAB over using Block Bot.

4.1.3 Research Questions

I explore the following research questions in this chapter:

1) How do perceptions of online harassment vary between users who subscribe to blocklists and users who are blocked on these lists? What behavior patterns do blocklist subscribers identify as instances of online harassment?

2) What motivates users to subscribe to anti-abuse blocklists and how do their experiences change after subscribing? What are the advantages and challenges of curating blocklists using human moderators? How do blocked users perceive the use of anti-abuse blocklists?

4.1.4 Contributions

In this chapter, I contribute, first, a rich description of Twitter blocklists - why they are needed, how they work, and their strengths and weaknesses in practice. Second, I contribute a detailed characterization of the problem of online harassment. I include the perspective of people accused of harassment, which is often omitted from discussions of this topic. As we will see, both those who suffer harassment and those who are accused of it are diverse groups. This diversity matters when we plan solutions. Further, I explore the idea that the flip side of harassment is understanding across differences - these problems are intertwined. Finally, I contribute a set of design challenges for HCI in addressing these issues.

The remainder of this chapter is organized as follows: I start by discussing background and related work that is important to understand the context of this study. Next, I present my methods of data collection and analysis. Following this, I discuss the qualitative analysis of my interviews and observations, focusing on my findings on online harassment and Twitter blocklists. In the final section of the chapter, I build on my findings to suggest a broad set of design opportunities that can help address online harassment on social media. I close with possible future directions of this study.

4.2 Related Work

4.2.1 Online Harassment on Twitter

Since its early days, Twitter has positioned itself as a platform for free speech. This supported the rapid gain in popularity of Twitter, and helped it play a critical role in many recent social movements – from the Arab Spring to Black Lives Matter (Lotan et al., 2011; De Choudhury et al., 2016). However, this maximalist approach to free speech also created conditions for online abuse on the platform (Warzel, 2016). A Buzzfeed report on Twitter noted that the platform treated online abuse as a "perpetual secondary internal priority" and allowed it to grow as a chronic problem over the last 10 years (Warzel, 2016). Twitter's unique design allows users who don't "follow" each other to interact, but it also makes it difficult to moderate content, because anyone can respond to any comment. This exacerbates the problem of abuse on the platform. This chapter provides insights into the nature of online harassment on Twitter.

4.2.2 Twitter Blocklists

Since the early days of the internet, social media sites have used blocking mechanisms to allow their users to filter the content they consume. Judith Donath describes how Usenet employed "killfiles," filters that allowed Usenet users to skip the unwanted postings (Donath, 1999). If a user put someone in their killfile, he stopped seeing any more of their postings. The use of killfiles was found to be effective in keeping the newsgroups readable. However, Donath also describes the resentment of users blocked on Usenet:

"To the person who has been killfiled, Usenet becomes a corridor of frustratingly shut doors: one can shout, but cannot be heard" (Donath, 1999).

Donath characterizes killfiles as “a good example of a social action that is poorly supported by the existing technology” (Donath, 1999). I assess in this chapter how far the technology has progressed to support the needs of social media users to ignore offenders by evaluating the use of contemporary blocking solutions, specifically blocklists on Twitter.

Stuart Geiger conducted a theoretical analysis of blocklists, and concluded that blocklists provide a concrete alternative to the default affordances of Twitter by facilitating a “bottom-up, decentralized, community-driven approach” for addressing online harassment (Geiger, 2016). Different individuals can have different perspectives on what online harassment entails, and where to draw the boundary between freedom of expression on the internet and online abuse. Geiger found that instead of a fixed Twitter-directed technological solution for addressing harassment, block-bots provide a social solution by allowing users with similar values to come together and engage in collective sensemaking. Geiger also notes that blocklists are “impactful in that they have provided a catalyst for the development of anti-harassment communities. These groups bring visibility to the issue and develop their own ideas about what kind of a networked public Twitter ought to be” (Geiger, 2016).

This chapter contributes to updating Geiger’s findings on blocklists (Geiger, 2016). I use empirical research methods to extend Geiger’s work by incorporating arguments made by users affected by blocklists.

4.3 Methods

This IRB-approved study adopts a mixed methods approach. I used the results of a network analysis on Twitter to select my sample of participants for semi-structured interviews. I focused on Good Game Auto Blocker (GGAB), a popular blocklist currently in use. To understand the motivations and experiences of blocklist users, I interviewed 14 users who subscribe to GGAB and triangulated my findings by interviewing 14 users who were blocked on GGAB and analyzing my participants’ posts on Twitter.

4.3.1 Participant Sampling

Sharan B. Merriam writes that "since qualitative inquiry seeks to understand the meaning of a phenomenon from the perspectives of the participants, it is important to select a sample from which the most can be learned" (Merriam, 2002). In selecting participants for this study, I used a purposive sampling approach. This approach advocates selecting participants who have rich information about issues of central importance to the research (Merriam, 2002). Inspired by Veldon and Legoze's study that combines network analytic approach with ethnographic field studies (Velden and Lagoze, 2013), I constructed a network of relevant users on Twitter and sampled the users most central to this network to recruit for interviewing. I expect that my centrality-based method of selecting interview participants helped me sample the dominant viewpoints of users.

As mentioned earlier, I interviewed two separate groups of participants: the first group is composed of users who were blocked on GGAB (hereafter referred to as UOB), and the second group contains people who subscribed to GGAB (hereafter referred to as SB). Next, I describe the details of how I sampled the participants for these two groups.

Selecting UOB Participants

I began by collecting the list of 9823 Twitter accounts blocked by GGAB. This list is publicly available on the BlockTogether website[8]. Next, I used the Twitter API[9] to retrieve the following information about each of the accounts on this list: (1) number of followers; (2) number of tweets issued; (3) date of account creation; (4) most recent tweet; (5) location; and (6) whether the account is verified.

I filtered out the accounts that were inactive (most recent tweet more than six months ago), created less than a year ago, verified (these are often accounts of brands and celebrities), located outside the US, had fewer than 20 followers, fewer than 100 tweets, or more

[8] The list of accounts blocked by GGAB is available at `http://tinyurl.com/ggautoblocker`.

[9] `https://dev.twitter.com/rest/reference/get/users/show`

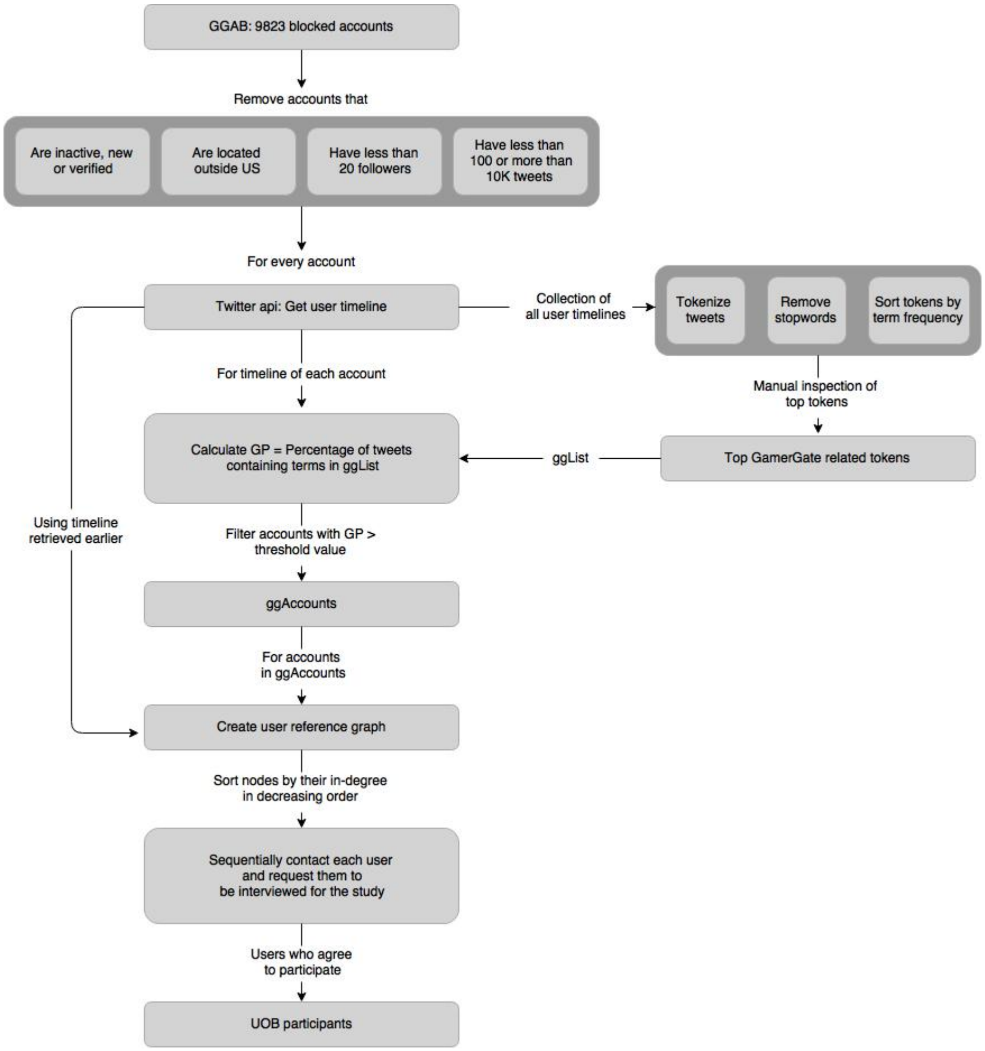

GGAB: 9823 blocked accounts
Remove accounts that
Are inactive, new or verified
Are located outside US
Have less than 20 followers
Have less than 100 or more than 10K tweets
For every account
Twitter api: Get user timeline
Collection of all user timelines
Tokenize tweets
Remove stopwords
Sort tokens by term frequency
Manual inspection of top tokens
Top GamerGate related tokens
ggList
For timeline of each account
Calculate GP = Percentage of tweets containing terms in ggList
Using timeline retrieved earlier
Filter accounts with GP > threshold value
ggAccounts
For accounts in ggAccounts
Create user reference graph
Sort nodes by their in-degree in decreasing order
Sequentially contact each user and request them to be interviewed for the study
Users who agree to participate
UOB participants

Figure 4.3: Steps taken to recruit UOB participants

than 10,000 tweets (these are often bot accounts). I call the list of remaining Twitter accounts, *blockedAccounts*.

I retrieved the Twitter timelines of *blockedAccounts,* and constructed a corpus of words that combined tweets from these timelines. I used a list of stopwords to filter out terms like ‘the’, ‘at’, ‘on’ from this corpus (Bird, Klein, and Loper, 2009). Next, I created a list, *tfList*, by arranging words from this corpus in decreasing order of their term frequency. I manually inspected the first 500 words in *tfList*, and extracted a list of terms related to Gamergate. I called this extracted list *ggList*, and it contains terms like ‘#Gamergate’ and

‘#sjwtears’.

For each account in *blockedAccounts*, I calculated *GP*, the proportion of all tweets containing any of the terms in *ggList*. I filtered out the accounts having *GP* below a fixed threshold level, *t = 0.3*. I also filtered out accounts not posting in English. I called the list of remaining accounts, *ggAccounts*.

The user reference graph

Next, I built a directed graph of accounts in ggAccounts, and found the accounts most central to this network. I treated each account as a node in this graph. If a Twitter account *a* mentioned[10] another account *b* (using @[handle]) in any of his posts, I added a directed edge from *a* to *b* in the graph.

I collected a ranked list of 100 nodes having the highest in-degrees in this network. These nodes represented accounts that are expected to be heavily invested in the Gamergate movement, and influential among the blocked accounts on Twitter. I then sequentially contacted these users on Twitter to recruit them for interviews. Figure 4.3 describes this process.

Selecting SB Participants

I used Twitter again to select interview participants who subscribed to GGAB. I collected all accounts that followed “@ggautoblocker,” the official Twitter account of GGAB. Following this, I used a process similar to the one in the previous section. I filtered irrelevant accounts and retrieved the Twitter timelines of the remaining accounts. As before, I created a Twitter network of these accounts using their mentions, and curated a ranked list of users central to the network.

I then contacted the users on this list by messaging them on Twitter. Since individuals who follow the GGAB Twitter account don’t necessarily subscribe to the GGAB blocklist,

[10] A mention is a post on Twitter that contains another user’s @username anywhere in the body of the post (Twitter, 2016b). Responses to another user’s tweet are also considered as mentions.

I asked all the potential interviewees whether they had subscribed to any blocklist. I only interviewed users who had subscribed to at least one blocklist.

4.3.2 Interviews

As discussed above, for each of the two groups, I invited the sampled users to participate in semi-structured interviews with me by contacting them on Twitter. About one in every five users I contacted agreed to do the interview with me. In all, I conducted 14 interviews with each of the two groups. Participation was voluntary, and no incentives were offered for participation.

The interviews began with general questions about which SNSs participants used, and their pros and cons. This provided necessary context to ask the participants about specific moderation problems and the use of Twitter blocklists. After developing some rapport with the participants, I asked them questions about their personal experiences and perceptions of online harassment. The interviews for the two groups followed different interview protocols. I conducted interviews over the phone, on Skype, and through chat, and each interview session lasted between 30-90 minutes. Some participants were contacted for brief follow-up interviews, for further clarification.

4.3.3 Participants

Most of the participants in my study reported being in their 20's and 30's. Among blocklist subscribers group, seven participants reported being male, four reported being female, two identified as transgender females, and one participant identified as non-binary. I read online postings and profile details of my participants on Twitter, and they indicated that some of the participants who identified as female in my interviews were also transgender. Among the participants who were put on blocklists, twelve identified as male and two as female.

Participants were self selected: I interviewed users who agreed to talk to me. Although most of my participants are from the US, I also had participants who live in Australia, UK,

Table 4.1: Blocklist Subscriber Participants

ID	AGE	GENDER	CISGENDER/ TRANSGENDER	OCCUPATION	COUNTRY
SB-01	36	Male	Cisgender	Web developer	USA
SB-02	21	Female	Transgender	Student	USA
SB-03	49	Male	Cisgender	Software engineer	USA
SB-04	24	Female	Cisgender	Student	USA
SB-05	24	Male	Cisgender	Courier	USA
SB-06	23	Male	Cisgender	Student	USA
SB-07	36	Male	Cisgender	Academic	Australia
SB-08	22	Female	Cisgender	Student	USA
SB-09	31	Female	Cisgender	Student	UK
SB-10	42	Male	Cisgender	IT consultant	UK
SB-11	41	Female	Cisgender	Physics instructor	USA
SB-12	26	Female	Transgender	Call center employee	USA
SB-13	27	Other	Not available	Student	Canada
SB-14	23	Male	Cisgender	Unemployed	Germany

Canada, Germany, Italy, Mexico and Netherlands. The interviewees included a blocklist creator and a blocklist moderator. Tables 4.1 and 4.2 provide some demographic information about my participants.

4.3.4 Analysis

I transcribed data from my interviews and read it multiple times. Next, I conducted an inductive analysis of these transcripts. I summarized my data with open codes on a line-by-line basis (Charmaz, 2006). I used the MAXQDA qualitative data analysis software (`http://www.maxqda.com`) to code my transcripts. Next, I conducted focused coding by identifying frequently occurring codes in my data and using them to form the higher-level descriptions. I then engaged in memo-writing and the continual comparison of codes and their associated data with one-another. I conducted iterative coding, interpretation, verification, and comparison through the course of the research. The comparisons led to the formation of axial codes that described seven overriding themes. In addition to the ones reported in the chapter, themes such as different perspectives on GamerGate move-

Table 4.2: Participants blocked by blocklists [a]

ID	AGE	GENDER	OCCUPATION	COUNTRY
UOB-01	38	Female	Childcare worker	USA
UOB-02	-	Male	Software developer	USA
UOB-03	28	Male	Consultant	USA
UOB-04	27	Male	PC repair	USA
UOB-05	36	Male	Medical professional	USA
UOB-06	33	Male	Game designer	Italy
UOB-07	20	Male	Student	UK
UOB-08	38	Male	Caregiver	UK
UOB-09	32	Male	Self-defense instructor	USA
UOB-10	32	Male	Appliance repairer	Mexico
UOB-11	33	Male	Game designer	USA
UOB-12	40	Female	Writer	UK
UOB-13	32	Male	Teacher	Netherlands
UOB-14	33	Male	PC repairer	USA

[a] All participants in this group are cisgender.

ment emerged but were excluded in further analysis. Finally, I established connections between my themes and these connections contributed to the descriptions of phenomena that I present in my findings (**Charmaz2006**).

4.3.5 Researcher Stance

The issues of online harassment and content moderation are sensitive, and as a researcher working in this space, I think it is important that I reflect on my position in this space. I identify online harassment as a systemic problem in the realm of the internet, and like many other systemic social issues, it disproportionately affects women and other marginalized groups. I share the conviction that urgent efforts need to be made to spread awareness about the extent and severity of the consequences of online harassment. While I respect every individual's right to freedom of speech, I also recognize that abuse and harassment should not be justified in the name of free speech. I further believe in the need for designers and researchers to create tools that provide victims of online harassment the necessary support and security. However, my stances have developed through the course of this thesis and I have come to see that such technologies can also carry the risk of falsely accusing individuals of harassment.

My goal in this chapter has been to investigate the effectiveness of one anti-abuse tool in addressing online harassment in a fair way. I have not evaluated the public or private communications of my participants with other users. Therefore, I am not in a position to pass judgment on whether online harassment occurred or did not occur in different contexts. However, my methods have allowed me to listen to my participants on both sides - those who used this tool to avoid being harassed as well as those who were identified as harassers and blocked by the tool - and understand their views on these complex issues. Therefore, I present my findings as subjective perspectives of my participants. With my analysis, I hope to inform the readers about the complexities and challenges of online moderation.

4.4 Findings : Online harassment

4.4.1 Different perceptions of online harassment

In this section, I discuss perceptions of online harassment from two sides: users who have subscribed to blocklists (SB users) and users who have been blocked on GGAB blocklist (UOB users). As I discussed in Chapter 2, online harassment is not defined specifically, and it can be difficult to distinguish harassers from non-harassers. By talking to both sides, a more nuanced narrative emerges than a simple contrast of good and bad actors.

Although the perceptions of online harassment generally vary with the users' overall experiences, many of my SB participants mentioned being disturbed, and in some cases, traumatized, by online abuse. Participant SB-11 said that she had to start taking anti-depressants in order to cope with harassment. Describing an incident in which she found doctored photos of herself, she said:

> *"They were extreme. Extreme. Violent, and things that just stuck in my head that I couldn't ... they weren't just gross. They were violent. I couldn't shake*

> *them. I had to take a break and they kept intrusively coming into my thoughts. It was really awful." – SB-11*

Participants characterized online harassment as acts ranging from someone posting a spoiler about the new Star Wars movie to someone sending them death threats. Four of my participants mentioned that someone had tried to get them fired from their job by contacting their place of work because of online disagreements.

Many users of the UOB group did not realize that online harassment can have serious consequences. A few UOB users said that they don't believe that online harassment is a legitimate problem because they can block or mute anyone that bothers them. Some participants proposed that online harassment shouldn't be taken too seriously. Participant UOB-06 said:

> *"It's certainly unpleasant but it has nothing to do with terms like "oppression" and "danger" that often get thrown around. I am an LGBT rights activist in Italy and I have met people that face some real oppression and danger in their lives. . . I find that every form of oppression that can be filtered out or avoided by closing a browser is more like an annoyance than a problem." – UOB-06*

Some participants told me that different users have different sensibilities, and often, individuals view the same discussion in very different ways because of the differences in their points of view, identities, or the issues that they are tuned into. They argued that these differences may contribute to some users perceiving that they are being harassed in situations that others consider as an expression of valid political speech. Furthermore, a few UOB participants noted that they have seen instances in which disagreements on an issue are deliberately portrayed as harassment by others. They considered such cases a strategy by their opponents to push a political agenda.

> *"There's a narrative that social justice ideologues try to push. That, for example, if you think the female pay gap is a myth (or even not as severe as [what]*

> *third-wave feminists say), [they claim] that you're sexist, misogynist, and if you're a woman, you have internalized misogyny." – UOB-11*

Some SB participants said that there are instances when the person doing the harassment doesn't fully realize the extent of impact of their acts upon the harassed users. Other SB users felt that harassers are often gullible individuals who are dis-informed about political issues. For example, Participant SB-01 believes that some GamerGate supporters become aggressive in their responses because of their basic misunderstandings about the nature of journalism.

> *"GamerGate supporters ally with hatemongers because they too feel like outsiders, like they're ignored, and joining a mob is the only way to get the narrative centered around them...Their anger is genuine, even if the narrative is false." – SB-01*

In contrast to the perspectives of SB participants, UOB participants felt that many online commenters overreact to trivial cases of perceived offenses. They also worried about the dangers of backlash against such overreactions. Participant UOB-02 felt that some users who promote socially progressive views and stand against online harassment actually hurt their own cause by dismissing anyone who disagrees with them on any issue:

> *"There are these kinds of social justice issues out there that are really, really important and that really address a lot of very real marginalization and just horrible things that are going on and I feel like to some extent, some of these bad actors in that space have kind of sullied the name of something that should be a lot more compassionate than it currently is." – UOB-02*

Some UOB participants felt that the media often highlights the online harassment of a few groups while ignoring similar abusive behavior against opposing groups. Participant UOB-04 expressed sympathy for the targets of harassment but felt that it was duplicitous of media and many people with authority to depict different instances of online harassment in ways that he considered biased:

> *"It is hypocritical to me, though, because while I often see one specific group of people and the issues they face being portrayed as harassment (and this group of people are often friends of the people doing the portrayal) - I see other groups of people who face similar hardships being written off and undermined by the same group." – UOB-04*

Participants noted that prominent perpetrators of harassment include groups ranging from GamerGate supporters and GamerGate opponents to radical feminist groups. Some participants pointed out that harassers also include trolls who conduct harassment as a kind of cultural performance art. This is similar to what Whitney Phillips found in her work on trolls in which she provides an empirical account of the identities, attitudes and practices of trolls, and their impact on the digital media environments (Phillips, 2015b). She argues that "the vast majority of trolling is explicitly dissociative. . . the mask of trolling safeguards trolls' personal attachments, thereby allowing the troll to focus solely on the extraction of lulz[11]" (Phillips, 2015b).

A few of my SB participants believe that, in contrast to trolls, there are harassers who develop an emotional investment in hurting their targets. For example, Participant SB-11 told me that harassers include users with serious psychological challenges who deal with their personal traumas by attacking others online. She argued that such harassers exhibit characteristics that are quite distinct from trolls, for example, they often don't have their identity anonymized on social media. She distinguished such users from trolls by saying:

> *"Some of the stuff they (trolls) said was kind of shock value. . . they weren't obsessed with me. They didn't have some sort of emotional investment in hurting*

[11] "lulz" is a corruption of lol (laugh out loud) that signifies unsympathetic laughter, especially one that is derived at someone else's expense (Phillips, 2015b).

> *me. When you find those people who are really invested in you personally, for whatever reason, God knows why, that's the scariest ass thing." – SB-11*

Some UOB participants complained that they were perceived as harassers by other users because of their mild association with controversial individuals on Twitter, and not because of their own activities. Participant UOB-01 told me that she was harassed by many users and was accused of being a "gender traitor" after she was put on GGAB blocklist. She questioned the decision-making process of GGAB moderators and felt that they did not have any qualms about blocking the users who don't harass:

> *"I've always tried to talk to them but they simply don't care if you're a good person. If you don't agree with their ideals, you're automatically the bad guy. . . I tried to get removed [from a blocklist] and was denied because I retweeted people like Totalbiscuit and Adam Baldwin. I was guilty by association." – UOB-01*

4.4.2 Tactics used by harassers

In this section, I discuss some behavior patterns and tactics that my participants identified as manifestations of online harassment. Table 4.3 lists these tactics among others and briefly defines them.

Subtle threats

Some participants argued that often, the perception is that online harassment is transparently malicious, involves violent threats, etc. but online harassment can manifest more subtly too. For example, Participant SB-01 said that he received messages from strangers, which indicated that they had gleaned a lot of personal information about him from his social media postings.

Table 4.3: Tactics used by harassers

Tactic	Description
Brigading	A large number of users, often those belonging to the same group, posting together on other online spaces in order to disrupt conversations.
Concern trolling	Visiting a site of an opposing ideology, and disrupting conversations or offering misleading advise in the guise of supporting that ideology.
Dogpiling	Many users posting messages addressed to a single individual. The intent of any sender may not be to perpetrate harassment, but it results in the targeted individual feeling vulnerable.
Dogwhistling	Using messages that sound innocuous to the general population, but have special meanings for certain groups. Such messages are used as a covert call to arms to target an individual or a group.
Doxing	Revealing someone's private information online with an intent to intimidate them or make them vulnerable to offline attacks.
Identity deception	Providing a false impression of one's own gender, race, etc. to gain advantage in online conversations.
Multiple SNSs	Using multiple social network sites to retrieve more information about the targets.
Sealioning	Politely but persistently trying to engage the target in a conversation. Such conversations are often characterized by asking the targets for evidence of their statements.
Sockpuppeting	Using an alternate account to post anonymously on social media. This is often done to feign a wider support of one's own postings.
Subtle threats	Using subtle hints to intimidate targets and make them aware that their personal information is exploitable.
Swarming	A group of users simultaneously attacking the same individual.
Swatting	Anonymously contacting and misleading law enforcement to arrive at the unsuspecting target's address.

> *"Some strategies I have noticed: [messages like] 'Paul [12], you work in tech in Portland, how can you be so ignorant about this issue?' Another common example was mentioning to me that I have kids or making comments on selfies I'd shared a few weeks ago." - SB-01*

Participants believe that such messages intend to make the recipients aware that the harassers have read through many of their posts, and that their personal information is exploitable.

Using multiple social networks

A few participants reported that some trolls had gone through their profiles on multiple social media sites in order to gain personal information about them.

> *"A few days ago, I had a men's right activist who I blocked on Twitter and then he went and found my Facebook account and sent me threats through Facebook." – SB-07*

Some participants also noted that a few troll groups organize harassment on other, more obscure websites, but carry it out on more popular social media platforms like Twitter, taking advantage of their ineffective moderation.

Dogpiling

Some participants said that sometimes, the harassment is because of the volume of tweets. They referred to such cases as 'dogpiling'. They felt that in such cases, the intent of many posters may be that they want to debate someone who they don't know, but whose post they come across. Participant SB-11 said that in such cases, "when you're getting like a hundred messages all wanting to debate you, then it feels like you're being overwhelmed." Other participants felt that such dogpiling occurs as a result of coordinated troll campaigns.

[12] Name changed to preserve anonymity

> *“The first mass contact would be like shining a spotlight on you, and then all that other stuff would happen - digging in, contacting your family, or your employers” – SB-02*

Some participants noted that dogpiling occurs on Twitter when one of the participants in a discussion has a large number of followers who interject themselves in the conversation.

> *“Somebody was bothering J.K. Rowling and saying rude things to her, and she responded with something like, “Yeah, but your screen name is stupid,” or whatever, right? Because J.K. Rowling has millions of followers, this person just got descended on.” – SB-02*

Participants said that in such incidents, the individual with many followers may or may not have the intention to dogpile a target.

Identity Deception

Some previous studies have noted that trolls and harassers engage in identity deception (Donath, 1999), and its varieties such as gender deception (Turkle, 2006) and age deception. My participants consider such behaviors as manifestations of online harassment.

A few participants told me that in some cases, harassers use identity deception to strengthen their argument in discussions. They present themselves as belonging to a minority group or an oppressed community that they don’t actually belong to. Participants felt that this is done so that users who wish to be sensitive to the opinions experienced by minority groups don’t contradict them.

> *“Those that come out and say I’m a 13-year-old trans-girl from this Christian conservative family and I’m having such a hard time, and you know what I think? Then they just go on some rant, strong argument rant about something to try to prove that all these social justice people are completely ridiculous and*

> *a lot of people go - I'm not going to call out a 13 year old girl because that doesn't feel right." - SB-11*

Participants noted that trolls also frequently use identity deception to harass others. For instance, some GamerGate supporters claimed that a few troll groups incited both GamerGate supporters as well as opponents by posing to be on the other side, and posting offensive messages.

Some participants said that many harassers and trolls create multiple accounts, and use them to overwhelm their targets. In some cases, multiple accounts are used to pretend that other users agree with and support their abusive responses to the target. Participant SB-10 said that when an account gets banned on Twitter, the abusers often quickly make a new 'sock' account, and resume attacking their targets (See "sockpuppeting," Table 4.3).

Brigading

Brigading refers to a concerted attack by one online group on another group, often using mass-commenting or down-voting. Some of my participants who actively use Reddit noted that on Reddit, brigading frequently occurs on subreddits with opposing ideologies. When someone posts on subreddit r_1 a link to a submission or comment *c* posted on a different subreddit r_2 with the knowledge that *c* would be unpopular on r_1, it has the effect of r_1 users down-voting *c* on r_2 and posting replies to *c* that are contrary to the values of r_2 users.

Some participants noted that on Twitter, brigading often occurs through malicious misappropriation of hashtags. They mentioned that trolls often "brigade" hashtags that are used by minorities, and disrupt their conversations.

> *"Many of them would flood tags like #ICantBreathe in support of Eric Garner and other victims of police violence with racism and gruesome images or pornography." - SB- 13*

Some users noted that brigading is sometimes used to spread misinformation. They also pointed out that feminist hashtags like #YesAllWomen, tags used by disability advocates, and tags that abuse victims use to share their stories are also frequently brigaded.

Sealioning

A few participants told me that harassers, particularly those who support GamerGate, engage in persistently but politely requesting evidence in a conversation, a behavior referred to as "sealioning." The name "sealioning" is derived from a comic in which a sea lion repeatedly tries to talk to a couple and annoys them (Malki, 2014). Sealioning is viewed by many users as intrusive attempts at engaging someone in a debate.

Doxing

Doxing refers to revealing someone's private information online. My participants told me that doxing often occurs on relatively obscure internet forums that are dedicated to doxing, but it also occurs on more popular social media platforms. A number of participants from both groups mentioned that they were doxed online or had witnessed doxing. Doxing a user using a pseudonym can be damaging to that user, because it may reveal information about that user that he may not be comfortable associating with publicly. Often, doxing involves revealing information such as social security number or residential address. In such cases, many fear that the doxed person's personal safety can be endangered.

Some participants expressed frustration that their reporting of doxing is met with the platforms' response that the doxed information is available through a search of their username on internet search engines, and therefore, it cannot be removed because it is public information. Others complained that the platforms respond to doxing only if the person being doxed is a public figure.

Other privacy intrusions identified by my participants include attempts to hack online accounts and calling employers to try to get the targets fired. They also noted that in many cases, threats of these actions are used to harass individuals.

4.4.3 Who is vulnerable to harassment?

Perceiving minority groups to be more vulnerable

Online harassment knows no boundaries, and virtually anyone can become a target. However, some of my SB participants insist that certain groups like the transgender community and Muslims are especially vulnerable to online harassment. Some participants noted that females, particularly women of color and feminists, are another vulnerable group. Many participants felt that the identity of harassed users plays an important role in how much and in what ways the harassment is conducted. Others pointed out that they enjoyed privileges due to their gender, race or nationality, and were not very vulnerable to harassment.

Believing that dependence on online communities for social support increases vulnerability

My transgender participants told me that the online transgender community on Twitter is very important to them, because it allows them to connect to individuals with similar experiences.

> *"I'm transgender, something it's taken me years to come to terms with. And then over the past couple of years, I've learned that there's a whole community of people a lot like me. There's a lot of bonding." – SB-12*

They felt that the dependence of transgender users on their online community as a primary means of social support makes them more vulnerable, because it is more difficult for them to leave the platform. Participant SB-12 mentioned that some of the harassment of transgender users comes from other users within their community.

> *"Trans people, especially trans women, are often harassed online. As much as it pains me to say it, we aren't always good to each other, either. I've seen trans women make death threats toward each other." – SB-12*

Believing that a decrease in anonymity increases vulnerability

Early research on online communities has shown that the relatively open nature of the information on SNSs and many users' lack of concern with privacy issues expose users to various physical (e.g., stalking) and cyber (e.g., identity theft) risks (Gross and Acquisti, 2005). Users who have taken great efforts to create an online presence on a platform find it difficult to leave the platform (Donath, 1999) if they face abuse, and are therefore vulnerable. Some of my interviewees also said that there is a connection between anonymity of a user and his or her vulnerability to harassment.

> *"I found that the less anonymous you are, the more cruel people can be." – UOB-01*

Some participants felt that revealing too much information about oneself or being too outspoken about sensitive issues on some platforms can be dangerous. However, they also found it difficult to determine exactly how much personal information is "too much" information in a given context.

4.4.4 Support of harassed users

In their study of Hollaback [13], a social movement organization whose mission is to end street harassment, Dimond et al. found that sharing and reading others' stories of how they define harassment and respond to it helped shift harassed individuals' "cognitive and emotional orientation towards their experiences" (Dimond et al., 2013). I found evidence of participants drawing comfort from sharing their experiences in my study too. My participants appreciated the support that they received from other users on the platform during episodes of online abuse. For example, Participant UOB-01 described how she drew comfort from sympathetic messages from GamerGate supporters:

[13] https://www.ihollaback.org

> *"Yes, I had random strangers tweet at me their support... Even when I had angered a few people by leaving GamerGate the way I did, I had many still in the movement show their support. Twitter has never been short of amazing people. It's just the louder, angrier voices carry more weight." - UOB-01*

Participant SB-11 discussed how many harassed users drew comfort from her volunteer work of blocking and reporting accounts who harass them online.

4.5 Findings : Twitter Blocklists

I discuss my findings on blocklists in this section. My discussion of different perceptions of online harassment in Section 4.4 provides context for understanding different views of the use of blocklists in this section.

4.5.1 Algorithmically curated versus socially curated blocklists

Some popular blocklists (e.g., GGAB) are algorithmically curated while others (e.g., Block Bot) are socially curated by a small group of volunteer users. When talking about the different blocklists that are popular on Twitter, Participant SB-11, a Block Bot moderator, drew distinctions between algorithmically and socially curated blocklists. She said that although algorithmically curated blocklists cannot make the complex decisions that humans can make via a socially curated blocklist, their decisions may be perceived by blocklist users as more objective, as they have predefined, fixed criteria. On the other hand, Participant SB-11 further argued, socially curated blocklists would tend to have fewer false positives [14] because they are curated by humans.

Block Bot, in particular, blocks a Twitter account when two different moderators agree that the account deserves to be blocked. Additionally, every Block Bot moderator has

[14] Throughout this chapter, I use 'true positives' and 'false positives' in the context of blocklists metaphorically in order to refer to accounts that most subscribers of the given blocklist would prefer to block and not block respectively. I do not make any claims about whether the 'true positives' are genuine *harassers*. Making such claims would require operationalizing what exactly counts as harassment in the analysis at hand, which is beyond the scope of this thesis.

the right to unblock any blocked account. Block Bot moderators believe that this reduces the probability of having false positives. Different users may have different views of who should and who should not be blocked, and therefore, it is difficult to predict how many Block Bot subscribers would agree with its moderators' decisions. I note however that many participants complained about the high number of false positives on GGAB, an algorithmically curated blocklist, but I didn't hear such complaints about Block Bot in my interviews.

> *"When Harper[15]'s GGAB was made, it was so broad that it had company accounts on there, e.g., KFC twitter [account], which follows back people who follow it." – SB-07*

4.5.2 Why do users subscribe to/avoid subscribing to anti-harassment blocklists?

When I asked SB participants how they came to start using blocklists, some users in the sample mentioned that they subscribed to blocklists after receiving targeted harassment. They pointed out that often, these harassers belong to the same group such as GamerGate or TERFs (trans-exclusionary radical feminists). Some users noted that they started using blocklists because reporting abuse to Twitter proved to be futile.

> *"It would have been early last year when I sent a tweet using the #GamerGate, and I compared GamerGate to men's rights activism...I received hundreds of tweets every hour, and I discovered that there were discussion boards where I've become the focus of discussion among people participating in GamerGate...That went on for about a week or so until I was able to install an auto blocker and basically just cut them out, so they couldn't continue the harassment." – SB-07*

[15] Randi Harper is a Software Engineer, based in Portland, Oregan. She was involved in the GamerGate controversy and has created a few open-source anti-harassment tools including GGAB (Patreon, 2017) .

> *"I eventually just grew tired of trying to either respond to these people and yeah, eventually, I just didn't want to talk to these people anymore, because it was a repeating pattern...They say the same offensive remarks, same insults, same aggressive behavior. It wasn't really changing anything when reporting them to Twitter. " – SB-14*

Other participants preemptively subscribed to the blocklists because they thought it would shelter them from seeing hate filled content. Some participants assumed that users who are on blocklists aimed at blocking a specific deviant behavior would be bigoted in other ways too.

> *"A lot of people are actually on multiple of these block lists because a lot of the anti feminist and the GamerGate people are one and the same. Also a lot of racists are also one and the same people. It's mostly a very similar kind of people." – SB-14*

Many users who subscribed to blocklists also promoted the use of blocklists on Twitter and other social media as a solution to mass harassment on Twitter. Participant SB-05 said that he followed many individuals who were harassed by the GamerGate movement, and who used and promoted GGAB. This convinced him to preemptively start using GGAB in 2014. Some users told me that the use of anti-abuse blocklists was so popular among pro social-justice users, that anyone who followed such users knew about blocklists.

> *"Nobody really questions the necessity of GGAB, since [Gamer]gators in your mentions are like an STD. You did something wrong, and you need to get rid of them." – SB-12*

Some participants took to using blocklists after they failed to find a common ground with groups of users with opposing views. Participant SB-12 mentioned that she used GGAB because after having a few discussions with GamerGate supporters, she realized that neither she nor the GamerGate supporters had anything to gain from listening to one another. Similarly, Participant SB-01 said:

> *"I was getting constant mentions from GamerGate accounts, and I was wasting tons of energy replying. When someone asked "what have you been up to?", all I could think of was: arguments with anonymous neofascists. I saw a few people discussing preemptive blocking and looked into it." – SB-01*

In addition to GGAB and Block Bot, some participants also use other anti-abuse blocklists. For example, Participant SB-08 uses a blocklist compiled by her and her friend in addition to GGAB. Participant SB-01 uses a blocklist manually curated by a widely-followed artist. Participant SB-14 maintains his own private blocklist and shares it with a few of his friends.

None of the UOB participants subscribe to any anti-abuse blocklists like GGAB and Block Bot, often citing the reason that they are opposed to avoiding discussions with those they disagree with. Some SB participants also deliberately avoided using blocklists despite claiming that they suffered online harassment. For example, Participant SB-03 described why he doesn't use any blocklists:

> *"I find it useful to follow some of the ringleaders of these brigades to keep track of them...A disadvantage of using blocklists is that you may miss early warnings, or actual engagement." – SB-03*

Participant SB-13 used three blocklists but he avoided using GGAB blocklist because he felt that he wasn't high profile enough to be a direct target of GamerGate supporters, and he could avoid becoming a target just by not using the hashtag #GamerGate.

4.5.3 How did user experience change after using anti-abuse blocklists?

Many SB users said that their Twitter experience improved after they started using anti-abuse blocklists. They told stories about how they stopped getting large numbers of unwanted notifications, and tended to get only genuine inquiries from strangers instead of abusive messages.

> *"It does quite a good job of putting up a firewall between me and sections of Twitter that I don't really want to have access to my content and I don't really want to engage in dialogue. By and large, I've found that it's a pretty effective way of taming Twitter." – SB-07*

> *"Twitter is a cleaner place when I go looking at things. I don't see the thousands of things from racist, bigoted people. It's just nicer for me to not have to see terrible stuff written on a daily basis, which is good." – SB-08*

Participant SB-10 felt that a few minority groups, for example, transgender communities, especially benefit by the use of anti-abuse blocklists:

> *"I certainly had a lot of transgender people say they wouldn't be on the platform if they didn't have The Block Bot blocking groups like trans-exclusionary radical feminists – TERFs." – SB-10*

Participant SB-02 said that she noticed a decrease in unwanted notifications in stages. When she first subscribed to a blocklist, her notifications decreased a lot and then leveled off. After some time, such notifications decreased further. I suspect that this decrease in notifications could have corresponded to the periodic blocking of new accounts by the blocklist.

Many participants found the use of blocklists liberating. Participant SB-11 said that the use of blocklists empowers individuals to set their own boundaries by providing them with effective tools to control the content they consume. None of the participants who used blocklists seemed to be bothered by whether blocklists blocked users who they would not have blocked. For example, Participant SB-09 said that she noticed a few accounts that were blocked because of the blocklist's false positives and she manually unblocked these accounts. However, the benefits of having fewer unwanted notifications and abusive messages far outweighed this cost for her. Many other participants shared a similar view:

> *“If I* really *want to know what blocked people are saying to me, I can go check from an incognito browser, but realistically, there are millions of people on Twitter. Saying I don’t want messages from less than 1% of them isn’t going to damage the range of people I can hear from.” – SB-01*

> *“It’s not 100% foolproof, but it’s very effective. Sometimes I find people that I like that happen to have blocked, so I unblock them. Sometimes it’s a little tedious, but it’s worth it.” – SB-06*

> *“It’s not a perfect solution, but no anti-harassment tool is going to be perfect. I’m much more worried about the impact of targeted harassment on me and my career than I am about the minor inconvenience that someone might face because they are unable to contact me on Twitter.” – SB-07*

Many participants who found blocklists useful still pointed out that they would much rather have Twitter address the problem of harassment effectively so they don’t have to use third-party tools like blocklists. Not all the participants consistently found blocklists dependable. Although Participant SB-09 found blocklists useful when she first subscribed, she feels differently about them now:

> *“I’m far from keen on them, now that there are better tools out there, and after someone with a large following used them to cut a lot of innocent people off from a number of communities on Twitter by sharing her personal blocklist and claiming it was mostly ‘MRAs [Men’s rights activists] and trolls”’ – SB-09*

Participant SB-04 felt that although using blocklists improved her Twitter experience slightly, she still suffered harassment.

> *“I still got harassed but my experience was probably slightly better because of it. It’s not like it kept them all from making new accounts.” – SB-04*

4.5.4 Challenges of social curation

Motivating moderators

Social curation is not a trivial activity. It requires many volunteer users contributing several hours each week and coordinating with one another to moderate sexist, racist and homophobic content. Reviewing rape and death threats, violent images, and aggressive threats over a long period can be psychologically damaging. Some media reports have described how regulating the internet can deeply affect moderators and even drive them to therapy (Ruiz, 2014; Wagner, 2012).

What then motivates the moderators of socially curated blocklists to continue blocking trolls and harassers for free? I asked a Block Bot moderator what drives her to continue moderating, and she replied:

> *"One thing that motivates me is that I know that we're providing service that people need to stay connected to others. Especially with the trans-community where maybe in your city, in your community, there's a handful of other trans-people. Otherwise, you aren't connected to people with similar experiences as you...a lot of people said [to me] point blank: If I didn't have your service, I couldn't be online...Now, I can, and so I keep [more] connected to the world more than I would be able to otherwise." – SB-11*

Guarding against rogue moderators

Beyond the problem of keeping volunteers motivated to continue moderating, socially curated blocklists have a number of challenges. Participant SB-11, a moderator for Block Bot, mentioned that in rare instances, one of the moderators of Block Bot acted irrationally, and blocked a number of people who didn't deserve to be blocked. Following this, the Block Bot team did some damage control, and put in technical fail safes. However, such incidents highlight the vulnerabilities that any socially curated blocking mechanisms can have.

Making decisions about perpetrators from vulnerable groups

Moderation decisions can be challenging for the moderators. When an account in question belongs to an abusive user from a vulnerable group, the decision of whether or not to block that account becomes difficult. For example, Participant SB-10 said:

> *"You might get an abusive member of the transgender community, and then the question is, do you block them and isolate them from their own community, given that the suicide rate in transgender people is actually very high. " – SB-10*

Moderators and users having different perspectives

Another challenge is that the moderation decisions of a socially curated blocklist are biased by the particular perspectives of the blockers. Everyone has different viewpoint and tolerance level, and what might offend one person may be perfectly reasonable to another. When users subscribe to the blocklist, they block accounts that are moderated through complicated decisions taken by the blocklist curators. This can be problematic because the users and the blocklist moderators may have very different definitions of what harassment entails. As Participant UOB-09 described, "You are guilty if they (blocklist moderators) say you are." Participant SB-11 explained:

> *"Most of the trans-people we've had as blockers are trans-women. They're going to have a certain perspective on what would be considered a blockable offense that's going to be slightly different than somebody else." – SB-11*

Other participants also worried about this problem, and felt that socially curated blocklists may be efficient only when they are curated to serve specific social groups constituting of individuals with similar ideas of what harassment means to them.

Resolving conflicts among moderators

Participant SB-11 told me that there are even differences among the Block Bot moderators on who they consider should be blocked. This can result in conflicts among the moderators.

> *"If somebody decides that this person shouldn't be on the [Block] bot, they just aren't. The only time that things have gotten really, really difficult is if there's somebody really close to my social circles who somebody feels strongly that their account should be placed on the Block Bot. That's happened a couple times and it was pretty terrible. We've avoided that sort of thing recently because you learn from when it happens." – SB-11*

Making trade-offs between being transparent and resisting attacks

The Block Bot moderators have made the list of blocked users publicly available on their website, www.theblockbot.com. They felt that having this list public has made the Block Bot more transparent as the potential subscribers can find out what they are signing up for by browsing the profiles of users who have been blocked. However, it has also made the moderators more vulnerable to attacks by users who are put on the blocklist. They said that they would rather ignore such users than annoy them or engage with them.

The Block Bot moderators save the tweets that led to each user being blocked and present them when any blocked user inquires them about the reason why he or she was blocked. They said that showing the posts that resulted in their blocking often results in convincing the blocked users to drop their appeal to be unblocked. Users who are put on the Block Bot are not informed that they have been blocked. Participant SB-11 described the implications of this decision:

> *"The pro was, of course, we're not interacting with them, we're not escalating anything. They might not even know the Block Bot exists and they've been on it for years. On the other hand, if I did make a mistake, somebody got on the*

> *Block Bot and they don't want to be on the Block Bot and they can't really talk to me about it if they don't know they're actually on it." – SB-11*

4.5.5 Perception that blocklists block too much/block unfairly

All the UOB participants felt that the currently popular blocklists block unfairly because they were surprised by finding themselves on blocklists and they did not feel that any of their actions warranted being blocked.

> *"If you suddenly get put on a list of "the worst harassers on Twitter" when you haven't said anything on the platform for years then you sort of want to know why." - UOB- 08*

Some participants felt that the criteria for curation of some blocklists - such as blocking all accounts who follow specific Twitter handles - was too crude to be considered reasonable.

Many participants worried about the personal biases of the blockers who compile socially curated blocklists or the developers who design or code algorithmically curated blocklists. A few SB participants also shared similar concerns and chose not to use some blocklists because they disagreed with the politics of individuals who managed those lists.

> *"GG block list is run by someone who had some skewed ideas of a few things that I happen to disagree with. Pretty bad stuff." – SB-06*

Some UOB users said that they couldn't access the pages of popular public figures, artists, etc. because they were using the blocklists the users were blocked on. A few UOB users claimed that they suffered professionally because of being put on blocklists.

> *"At a certain point, the International Game Developers Association sponsored the [gg]autoblock[er] claiming it was a way to block the worst harassers of Twitter. Beside not being a harasser, I am a game designer so an association that is supposed to protect me was accusing me instead." - UOB- 06*

Participant UOB-10 said that many new blocklists copy the accounts already put on popular blocklists. This leads to many users being blocked by accounts who wouldn't block them otherwise. A single individual's personal dislike and subsequent blocking of a user can snowball and end up excluding that user from many important groups. I also found evidence of this phenomenon outside my interviews. For example, one blogger complained that an influential Twitter user put many transgender users on her personal blocklist over minor disagreements: *"When someone in a position of trust and power blocks many marginalized people over minor disagreement, then it disseminates distrust and removes avenues of communication for those marginalized people...All those people are now blocked by all the tech contacts, feminists, celebs that sign up to her list, as well as anyone signed up to their block list"* (Sjwomble, 2016).

4.5.6 Feelings about being put on blocklists

When I asked UOB participants how they felt about being put on blocklists, their reactions ran the gamut from indifference and mild annoyance to disgust. Some participants did not feel very strongly about being put on blocklists. Others felt a little irritated that some users would consider them "horrible" without asking for any proof just because they are put on the blocklists. Still others felt okay with it because they believed that any Twitter user had the right to block whoever they want.

> *"We just sort of laughed at it and shook our heads since it seemed like a dumb thing to waste time on." - UOB-14*

Participant UOB-01 said that she was put on a blocklist just because of being part of GamerGate, and her efforts to be put off the list were futile.

> *"I never tweeted anything mean at any one. I always said harassment was wrong. It got me nowhere." - UOB-01*

Many UOB users described similar incidents of being put on a blocklist unfairly. They felt that they were victimized just for having wrong associations on Twitter, and that they did not deserve to be perceived as harassers.

A Block Bot moderator told me that some users got really upset about being put on Block Bot, because they thought that the blocklist was making an incorrect claim about the kind of person they are. Some UOB users criticized people who subscribed to blocklists. They characterized the blocklist subscribers as individuals who are not open to challenging their own conceptions or questioning themselves.

> *"If someone needs to insulate themselves via lists someone else created, they have bigger problems than getting offended by what I have to say." - UOB-12*

A few participants felt that it is unfair to be blocked for disagreement on just one topic of discussion, for example, their support of GamerGate. Participant UOB-06 said that it is cowardly of blockers to block the accused who then have no means of responding to the blockers' allegations.

4.5.7 Appeals procedure

Some SB participants said that they did not feel very strongly about the necessity of a fair appeals process.

> *"Being blocked on Twitter is not a legal issue, it's not a censorship issue, it's not a human rights issue...If there is a way for people to appeal, that's great. I don't actually think it's compulsory, to be honest." - SB-07*

Many UOB participants expressed an indifference about using an appeals process. For example, Participant UOB-12 said that it never occurred to him to try to get himself off the blocklists.

> *"People aren't obliged to speak to me. It's still a (fairly) free internet." - UOB-12*

Some UOB users were not aware of the existence of an appeals procedure that they could use to get off the blocklist. A few participants mentioned that they were discouraged from appealing to get off the blocklists because they had seen discussions about many cases of such requests by other blocked users proving futile. Participants UOB-01 and UOB-13 said that they sent messages to get themselves off the blocklists, but their requests were denied or they never received a response. This further contributed to their negative views of the use of blocklists.

> *"If you look at the behavior of the people who control these things, I think you'd have to be incredibly optimistic to expect a response, unless you are someone famous or rich or whatever." - UOB-13*

Many users said that they considered appealing to get off a blocklist to be too much of a bother and not worth the effort required.

> *"I don't really need an appeals process; again, that would be too much work. It's only if everybody I started following would block me because they were using the same block list, then I would go to it because I'm like, I'm not that bad." - SB-08*

This suggests that the existence of a fair appeals procedure may become more critical to those who are blocked if the number of users subscribing to a blocklist increases or if a blocking contagion occurs.

4.6 Discussion

Social media allows users having different experiences, ideologies, and political opinions to interact with one another. Twitter, in particular, as a result of its open design, allows users to find conversations on diverse topics, and respond to any posts directly. This provides an extraordinary opportunity for constructive discussions, understanding different perspectives, and discovering bipartisan solutions to complex societal problems. Unfortunately, in many instances, the interaction of users with opposing viewpoints results in aggressive behavior. In this section, I use the findings from this study and expand on previous work to propose tools and interventions that designers and policy makers should consider in order to help cultivate civil discussions, as well as reduce instances of and mitigate harm from online harassment.

4.6.1 Focusing on vulnerable groups

Some users do not realize that their experiences may be quite different from other individuals. My findings in this study show that many users don't grasp the emotional toll that their facetious remarks can exert on other users (Section 4.4.1). As Whitney Phillips describes, "even the most ephemeral antagonistic behaviors can be devastating to the target, and can linger in a person's mind long after the computer is powered down. This is especially true if the target is a member of a marginalized or otherwise underrepresented population, whose previous experience(s) of abuse or prejudice may trigger strong negative emotions when confronted with nasty online commentary" (Phillips, 2015b).

My findings also suggest that the identity of harassed users plays a role in making them vulnerable to online harassment (Section 4.4.3). This provides the following opportunities for researchers:

Study oppressed groups and talk to them.

Efforts to understand the specific needs of each oppressed group on an SNS, for example, through surveys or interviews of individuals from the group, can inform the modification of existing moderation mechanisms so as to better serve that group. Researchers can study the online activities of specific oppressed groups, and characterize the posting activities that lead to unusually high abusive responses. These findings can then be used to inform the group of the type of postings that have inadvertently invited abusive responses in the past. This knowledge may help such individuals make informed decisions on whether to engage themselves on certain topics.

In an ideal world, everyone would always be able to speak their mind without fear of harassment. It's important for system designers to work to achieve that goal. It would be unfortunate if marginalized groups learned to self censor. In the real world, however, consequences of certain kinds of speech can have a negative impact on individuals. Until the fundamental conditions that make the internet so conducive to harassment are ameliorated to some degree, it is strategic for individuals to at least be aware that something they are about to post has a high likelihood of provoking harassment. If we could create tools to alert individuals to that possibility, they could make a more informed choice about whether the benefits of particular speech outweigh the risks. Such a tool ideally could give people guidance on how to express the same ideas in ways that are less likely to attract abuse, or better still in ways that are more likely to be truly heard by the intended audience. I imagine such a tool could be of great use to individuals on both sides in this study.

Develop tools and systems that serve the special needs of vulnerable groups.

Language and actions that are abusive to a particular vulnerable group, for example, transgender users, may not be offensive to other users. Therefore, researchers and designers may need to develop specialized tools and systems for each group. User studies employing the individuals from the target group as participants who use these systems can be deployed to evaluate the effectiveness of such systems.

4.6.2 Designing support systems for harassed users

Platforms can also design to support users who suffer online harassment. Aggressive behaviors can be diffused by providing users with an alarm functionality that alerts their friends of their need for help. Support systems can also help users who have suffered similar abuse connect to one another, and share their experiences. My findings indicate that harassed users value messages of support during episodes of online abuse, even from strangers (Section 4.4.4). Therefore, systems that allow targets of harassment to receive support messages from other harassed users can help them cope with online harassment. In their study of Heartmob[16], Blackwell et al. argued that public demonstrations of support not only provide validation for targets of harassment, but also create powerful descriptive norms that help other users determine what behaviors are and are not appropriate in an online community (Blackwell et al., 2018a). I note that although support systems can provide critical support to harassed users, they are vulnerable to attacks by trolls and they need to be carefully designed to ensure that they are not misused.

Some harassed users may not be aware of how to use the tools on SNSs available to them. Platforms should provide guidelines and tutorials to their users on how to safeguard their privacy and use the anti-abuse tools available on the site. SNSs should also promote online resources like HeartMob so that the targets of online abuse can get information on supportive organizations and other helpful resources. Such measures would indicate to the harassed users that the platform is committed to addressing abusive behavior, and encourage them to not leave the site.

4.6.3 Improving blocking mechanisms

There are a number of ways that blocking mechanisms can be redesigned so that they better serve the needs of different user groups. My findings suggest that there is a need for decentralized blocking mechanisms like Twitter blocklists that operate separately from the centralized moderation provided by Twitter. However, certain measures need to be taken to ensure that these lists block fairly as well as serve their subscribers appropriately.

Using hybrid blocklists

Creating hybrid blocklists – lists that combine algorithmic and social curation (Section 4.5.1) – can be a promising approach. Such lists could rely on carefully constructed algo-

[16] `https://iheartmob.org`

rithms that surface offensive content and categorize it based on severity. Posters of blatantly abusive content can be blocked directly. For postings that are flagged by such algorithms as possibly abusive, human moderators can examine them and decide whether the posters should be blocked. These blocklists should also have sufficient fail-safe mechanisms built into them so that the actions of an intruder or a rogue moderator may be quickly reverted (Section 4.5.4).

My findings indicate that some users found themselves on popular blocklists because of a tenuous connection with controversial individuals on Twitter (Section 4.4.1). Human moderators can ensure that individuals are not blocked for trivial reasons like following an abusive user. They may also consider muting certain individuals instead of blocking them so as to avoid punishing certain actions disproportionately.

Such a hybrid mechanism would make the curation of blocklists more objective and efficient as well as ameliorate the risks of having a large number of false positives (Section 4.5.5). This mechanism could also be adapted to improve the accuracy of blocklists that are curated for purposes other than addressing harassment, e.g., spam blocklists.

Making blocklists more transparent

My findings in Section 4.5.5 show how branding a blocklist as a list of harassers can be dangerous, particularly if the list contains many false positives. The blocklist owners have a responsibility to communicate to their subscribers that some users may mistakenly be on the list. They should also make efforts to ensure that the users on the list are not discriminated against. Participant SB-11 described a few such efforts she made for Block Bot:

> *"Back in the day... the way that things were written out on their [Block Bot's] website were more blunt... I changed a lot of the wording so that it was a little less harsh. That seemed to help people not be as upset about it. We've always been a little irreverent about complaints because all it is is blocking someone, we're not saying that you're a terrible person or that whatever you*

> *do on Twitter rises to some legal definition of harassment or anything like that."- SB-11*

This suggests that clarifying the purpose of the blocklist can help de-escalate rancor from the blocked users. The blocklist administrators can also choose to explicitly discourage discrimination against blocked users for any purpose outside Twitter.

Different user groups have different definitions of harassment and distinct moderation needs (Section 4.5.4), and therefore they may need to subscribe to different blocklists. Therefore, multiple instances of blocklists should be constructed. Each such instance should clearly state its purpose, and its moderators should be aware of and have the capability to address the needs of its subscribers. Moderators should be encouraged to reveal aspects of their identities and experiences that shape how they moderate. This would allow subscribers to take into account the biases of the moderators before subscribing to any blocklist. I found that many users who were put on blocklists were frustrated because they did not know the reason for their being put on the list (Section 4.5.6). I posit that blocklists should record the reason why each account is blocked, the moderator who blocked it, and other relevant metadata. Providing information about the reason for being put on the blocklist when requested may encourage the acceptance of blocklists among many users.

Designing to avoid blocking contagion

In her book on community self-regulation, Ostrom writes that "graduated punishments ranging from insignificant fines all the way to banishment, applied in settings in which the sanctioners know a great deal about the personal circumstances of the other appropriators and the potential harm that could be created by excessive sanctions, may be far more effective than a major fine imposed on a first offender" (Ostrom, 1990). Drawing from Ostrom's work, Kiesler et al recommend using graduated sanctions to increase the legitimacy and effectiveness of sanctions in online communities (Kiesler, Kraut, and Resnick, 2012). They argue that "lighter sanctions mitigate the ill effects from inevitable mistakes in categorization" and "stronger sanctions are perceived as more legitimate when applied only after lighter sanctions have proven ineffective." In a similar vein, Forte and Bruckman found that in order to maintain local standards of content production, Wikipedia uses a series of graduated sanctions when behavior-related policy is broken – beginning with the posting of warnings and leading to banning from the site (Forte and Bruckman, 2008).

I discussed in Section 4.5.5 that blocking contagion could be a serious consequence of the popular use of Twitter blocklists. To prevent this problem, platforms like Block Together can draw lessons of graduated sanctions from the research described above and discourage the permanent blocking of blocked accounts. Instead, they can consider enforcing the blocks only for a limited time interval initially, and escalate sanctions if repeated misbehavior occurs, for example, by increasing the time interval for which the offending user is blocked.

Blocking mechanisms can also consider discouraging the outright copying of blocked accounts for creation of new blocklists. They can make such copying contingent upon the permission provided by some central moderators. A central supervision of blocklists that are currently in use, and a regular evaluation of whether they serve their stated purpose, would guard against misappropriation of blocklists.

Improving appeals procedure

I discussed in Section 4.5.7 that a dissatisfactory appeals procedure delegitimized the use of blocklists for many participants. The process of appealing to get oneself off any blocklist should be made more intuitive and efficient. Timely and appropriate responses to such appeals, along with an effort to spread awareness about the damaging effects of online abuse, would help such blocking mechanisms gain broader popularity on the site. I acknowledge that responding to appeals is expensive for the blocklist administrators. Therefore, I recommend focusing on having fewer false positives, and automating the process of responding to certain types of appeals.

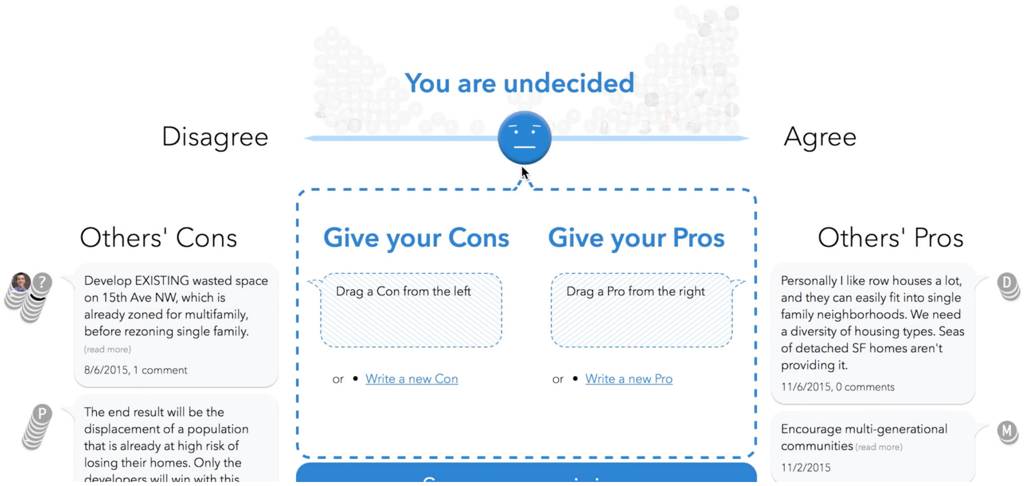
4.6.4 Building "understanding mechanisms"

My findings indicate that differences in identities, perspectives and sensibilities often contribute to situtations where some users perceive that they are being harassed and other users see it as mere disagreements (Section 4.4.1). Additionally, (mis)interpreting the words of the opposite side in a negative light and reacting inordinately over incidents of minor disagreements create further rifts and preclude productive discourse. Differences in behavioral and cultural norms across different user groups further escalate such situations. Furthermore, as Van Alstyne and Brynjolfsson warned in their study on "cyberbalkanization," if people spend more time on special interests and screen out less preferred content, it can "fragment society and balkanize interactions" (Alstyne and Brynjolfsson, 1996).

To address these challenges, designers need to focus on creating "understanding mechanisms." Tools that emphasize similarities between individuals could help them to understand one another and find common ground. Design solutions that allow users with different ideologies to interact without fear of being abused could foster productive discussions. There is a growing body of literature on modeling argumentation for the social semantic web (Schneider, Groza, and Passant, 2013). Designers can draw from the theoretical models and social web tools that argumentation researchers have proposed to implement mechanisms that facilitate constructive discussions.

Consider an open-source deliberation platform developed at the University of Washington, ConsiderIt (Kriplean et al., 2012), that powers the Living Voters' Guide[17]. This platform invites users to think about the tradeoffs of a proposed action by creating a pro/con list (Figure 4.4). This list creation is augmented by allowing users to include into their own list the points that have already been contributed by others. This process allows users to gain insights into the considerations of people with different perspectives and identify unexpected common ground (Kriplean et al., 2012). Additionally, the platform's focus on personal deliberation, as opposed to direct discussion with others, reduces the opportunities

[17] `https://livingvotersguide.org`



Develop EXISTING wasted space on 15th Ave NW, which is already zoned for multifamily, before rezoning single family. (read more)
8/6/2015, 1 comment

The end result will be the displacement of a population that is already at high risk of losing their homes. Only the developers will win with this

Personally I like row houses a lot, and they can easily fit into single family neighborhoods. We need a diversity of housing types. Seas of detached SF homes aren't providing it.
11/6/2015, 0 comments

Encourage multi-generational communities (read more)
11/2/2015

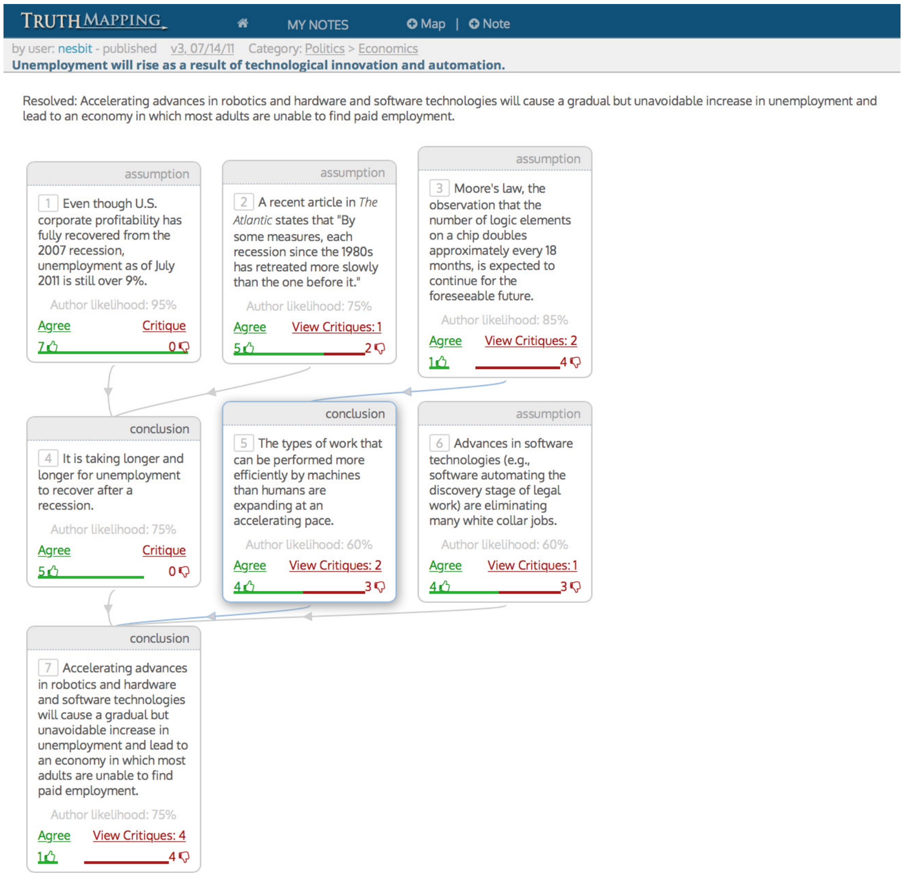
Figure 4.4: ConsiderIt allowing a user to compile pro and con lists for this proposal for Seattle city: "Increase the diversity of housing types in lower density residential zones."

for conflicts (Kriplean et al., 2012).

TruthMapping [18] is another online deliberation tool that allows users to collect and organize ideas, constructively test those ideas, and promote reasoning-based discourse. This tool structures conversations using argument maps, critiques and rebuttals (Figure 4.5). It invites users to break down a topic into its component parts – assumptions and conclusions – and create a node for each part, so that the hidden assumptions are made explicit. All critiques are directed against specific nodes so that any attempts at digression are apparent. Only the original arguer can modify the map but any user can provide their feedback by adding a critique to any assumption or conclusion or by responding to a previously posted critique with a rebuttal. As shown in Figure 5, TruthMapping also shows how many users agree or disagree with each node.

Although designs like ConsiderIt and TruthMapping offer innovative solutions to facilitating constructive deliberation, they assume that the users are participating in good faith, and are willing to devote their time to review previously posted content and submit productive accessions. These assumptions may not be true for many participants on social media sites. Therefore, designing "understanding mechanisms" for SNSs is a considerably hard

[18] https://www.truthmapping.com



by user: nesbit - published v3, 07/14/11 Category: Politics > Economics

Unemployment will rise as a result of technological innovation and automation.

Resolved: Accelerating advances in robotics and hardware and software technologies will cause a gradual but unavoidable increase in unemployment and lead to an economy in which most adults are unable to find paid employment.

assumption

1 Even though U.S. corporate profitability has fully recovered from the 2007 recession, unemployment as of July 2011 is still over 9%.

Author likelihood: 95%

Agree Critique

7 0

assumption

2 A recent article in *The Atlantic* states that "By some measures, each recession since the 1980s has retreated more slowly than the one before it."

Author likelihood: 75%

Agree View Critiques: 1

5 2

assumption

3 Moore's law, the observation that the number of logic elements on a chip doubles approximately every 18 months, is expected to continue for the foreseeable future.

Author likelihood: 85%

Agree View Critiques: 2

1 4

conclusion

4 It is taking longer and longer for unemployment to recover after a recession.

Author likelihood: 75%

Agree Critique

5 0

conclusion

5 The types of work that can be performed more efficiently by machines than humans are expanding at an accelerating pace.

Author likelihood: 60%

Agree View Critiques: 2

4 3

assumption

6 Advances in software technologies (e.g., software automating the discovery stage of legal work) are eliminating many white collar jobs.

Author likelihood: 60%

Agree View Critiques: 1

4 3

conclusion

7 Accelerating advances in robotics and hardware and software technologies will cause a gradual but unavoidable increase in unemployment and lead to an economy in which most adults are unable to find paid employment.

Author likelihood: 75%

Agree View Critiques: 4

1 4

Figure 4.5: TruthMapping allows users to construct an argument by laying out assumptions and conclusions.

problem, and there is a lot of potential for researchers to experiment with creative solutions in this space.

4.7 Conclusion

Online harassment is a multi-faceted problem with no easy solutions. Social media websites are persistently squeezed between charges of indifference to harassment and suppression of free speech (Auerbach, 2016). I believe it is an important and difficult challenge to design technical features of SNSs and seed their social practices in a way that promotes constructive discussions and discourages abusive behavior.

The emergence of third-party, open-source moderation mechanisms like blocklists introduces an innovative alternative to traditional centralized and distributed moderation systems. In this chapter, I focused on studying the effects of using blocklists - on those who used them and those who were blocked on them. I also used blocklists as a vehicle to investigate the broader issue of online harassment.

This chapter does not investigate all the possible forms and aspects of online harassment. Participants in the study were strategically recruited in ways that ensured awareness of and experience with these issues on Twitter. Other methods of recruiting may reveal other, perhaps more commonplace, experiences of average users with undesirable content and moderation. Researchers may also recruit users from specific vulnerable groups to understand their particular experiences and needs. Do these groups need moderation tools that serve their special needs? Can we design to detect distinctive harassment strategies such as dogpiling and brigading? Can we construct tools to combat these strategies that are not vulnerable to being abused? These are questions that are important to consider in future investigations of online harassment.

In the interim, by describing the experiences of users affected by blocklists on Twitter, I see concrete examples of the gap between the needs of users and the affordances provided by default and third-party moderation mechanisms on social media. If we hope to create scientifically informed guidelines for designers to follow, more work is needed that tests innovative design ideas for improved moderation in lab and field experiments.

CHAPTER 5
HUMAN-MACHINE COLLABORATION FOR CONTENT MODERATION: THE CASE OF REDDIT AUTOMODERATOR

5.1 Introduction

Automated tools are increasingly playing an important role in the regulation of posts across different social media sites. For example, many social media websites are using machine learning tools to identify images that violate copyright law and remove them (Roberts, 2014). As these tools become more and more sophisticated, their role in the enactment of content regulation will likely grow over time. It is also expected that future regulatory mandates will further heighten the need for automated tools because government regulators are increasingly expecting platforms to quickly remove hate speech and other illegal content (West, 2018). Therefore, it is critical that we understand the adoption and use of these tools in current moderation systems.

In this chapter, I discuss the use of automated tools for moderation on Reddit[1]. Reddit adopts a "community-reliant approach" (Caplan, 2018) to content moderation. That is, it is divided into thousands of independent communities, each having its own set of volunteer moderators, posting guidelines and regulation tools. While scholarly accounts of regulation mechanisms on discussion sites like Reddit have usually focused on how moderators collaborate with one another, create new rules, and interact with community members to enact efficient and acceptable content curation (Diakopoulos and Naaman, 2011; Lampe and Resnick, 2004; Matias, 2016c; McGillicuddy, Bernard, and Cranefield, 2016), relatively little research has focused on how moderators use automated mechanisms. Although some

[1] Findings from this study were published in 2019 in the ACM Transactions on Computer-Human Interaction (TOCHI) journal (Jhaver et al., 2019a). Iris Birman assisted me and my PhD advisors, Amy Bruckman and Eric Gilbert, guided me on this work.

scholars have recognized the importance of automated regulation tools (Roberts, 2014; West, 2018), at present, we lack a clear understanding of content regulation as a relational process that incorporates both automated tools and human labor.

In addition to filling this salient gap in research, studying automated moderation tools on Reddit is particularly important because they form a critical component of the Reddit regulation system and they perform large proportions of all regulation actions (Chapter 7). It is necessary to examine the sociotechnical practices of how human workers configure and use automated tools so that we can identify the structural challenges of current regulation systems. At a broader level, it is crucial that we understand the moderation apparatus that social media companies have built over the past decade if we hope to provide practical solutions to the hard questions being asked right now about how regulation systems can distinguish freedom of expression from online harassment (Chapter 4), or how they can make decisions on content that may be too graphic but is newsworthy, educational, or historically relevant (Gillespie, 2018b).

Unfortunately, given the black-boxed nature of many social media platforms, the process of content regulation remains opaque on many sites. That is, it is hard for outsiders to infer behind-the-scenes operations that guide it. Reddit, however, provides an excellent opportunity to study the sociotechnical details of how automation affects regulation processes because its moderators are *not* paid employees bound by legal obligations to conceal the details of their work but are volunteer users. Taking advantage of this opportunity, I conducted interviews with sixteen Reddit moderators, and analyzed the ways in which the use of automated tools reshapes how Reddit moderators conduct their work. In doing so, I offer insights into the tradeoffs that arise because of the use of automated tools, the tensions of redistributing work between automated tools and human workers, and the ways in which platforms can better prepare themselves to adopt these tools.

I explore the following research questions in this chapter:

1. How are automated tools used to help enact content regulation on Reddit?

2. How does the use of automated tools affect the sociotechnical process of regulating Reddit content?

3. What are the benefits and challenges of using automated tools for content regulation on Reddit?

While preparing for this study, I found that Reddit has an open-access API (Application Programming Interface) that allows bot developers to build, deploy and test automated regulation solutions at the community level. Access to this API has encouraged the creation and implementation of a variety of creative automated solutions that address the unique regulation requirements of different Reddit communities. I focused on one of the most popular automated tools, called Automoderator (or Automod), that is now offered to all Reddit moderators. Automod allows moderators to configure syntactic rules in YAML format (Figure 5.1) so that these rules make moderation decisions based on the configured criteria. I show that Automod not only reduces the time-consuming work and emotional labor required of human moderators by removing large volumes of inappropriate content, it also serves an educational role for end-users by providing explanations for content removals.

Despite the many benefits of using Automod, its use also presents certain challenges. Prior CSCW research has established that a fundamental social-technical gap exists between how individuals manage information in everyday social situations versus how this is done explicitly through the use of technology. Often, technical systems fail to provide the flexibility or ambiguity that is inherent in normal social conditions (Ackerman, 2000). In line with this, my findings reveal the deficiencies of Automod in making decisions that require it to be attuned to the sensitivities in cultural context or to the differences in linguistic cues.

Building on this case study of Reddit Automod, I provide insights into the challenges that community managers can expect to face as they adopt novel automated solutions to help regulate the postings of their users. For example, they may have a reduced level of control over how the regulation system works — as moderators reduce the number of posts that they manually review and delegate to automated tools, these tools may make mistakes that could have been avoided owing to their limitations of evaluating the contextual details. Moreover, moderators may not be able to understand the reasons behind some actions taken by automated tools. Another possible challenge is that moderators may have to make decisions about the levels of transparency they show in the operation of automated tools — if they are too transparent about how these tools are configured, these tools may be exploited by bad actors.

In addition to these challenges, I also highlight how the use of automated tools may affect how moderators design community guidelines. I found that Reddit moderators sometimes create posting guidelines that play to the strengths of Automod so as to make the work of moderation easier. For example, guidelines like "describe the image you're posting" provide additional material for Automod to catch. However, complying with such guidelines may increase the amount of work that end-users have to perform. Therefore, using automated tools may affect not only the moderators but also the other stakeholders in content regulation systems.

There is a growing enthusiasm among many companies hosting user contributions to use machine learning and deep-learning-based tools to implement content regulation and relieve human moderators (Caplan, 2018; Madrigal, 2018; Ong, 2018). While the accuracy of such tools has risen for many kinds of moderation tasks, the tools often can't explain their decisions, which makes mixed-initiative human-machine solutions challenging to design. Human-understandable ML is an active area of research (Lakkaraju, Bach, and Leskovec, 2016). Yet, as we will see, Automod does not rely on machine learning techniques but it rather uses simple rules and regular-expression matching which can be understood by technically savvy human moderators. I found that moderators self-assess their skills at configuring Automod, practice care when editing Automod rules, and coordinate with other moderators through external channels like Slack to resolve disagreements. Thus, Reddit moderation is an intriguing example of human-machine partnership on a complex task that requires both rote work and nuanced judgment.

Note that the design claims presented in this and the next chapter are not prescriptive rules that a designer should follow blindly. Although they provide empirically grounded claims about the likely effects of design alternatives *on average*, designers should evaluate whether there are any extenuating circumstances that may produce unexpected results in their particular situations. Further, designers typically make multiple design choices at the same time and a given set of design choices may complement or contradict one another in producing desirable outcomes (Kraut and Resnick, 2012). Therefore, the design suggestions presented in these chapters should be viewed as guidance on the likely impact of design changes.

I organize the remainder of this chapter as follows: I first present my methods of data collection and analysis, and describe my participant sample. Next, I discuss my findings on the development and use of automated moderation tools, focusing on Reddit Automod as a case study. I then discuss the limitations of currently used automated techniques and the challenges they pose for Reddit moderators, emphasizing insights that may be useful for other platforms as they adopt automated regulation tools. I conclude with highlighting the takeaways of this research for different stakeholders.

5.2 Methods

This study was approved by the Georgia Institute of Technology IRB. The study included 16 in-depth, semi-structured interviews with Reddit moderators. Next, I provide the details of my study design.

5.2.1 Selection of Subreddits

Prior research suggests that as communities grow, human-only moderation becomes increasingly unfeasible, and the dependency on automated tools increases (Binns et al., 2017;

Table 5.1: Activity level, date of creation and number of moderators for sampled subreddits

Subreddit	Total comments	Creation date	# Moderators
oddlysatisfying	180,978	May 15, 2013	22
politics	14,391,594	Aug 06, 2007	37
explainlikeimfive	964,821	Jul 28, 2011	38
space	795,186	Jan 26, 2008	23
photoshopbattles	300,369	Jan 19, 2012	23

Gillespie, 2017a; Roberts, 2014). My long-term experiences as a moderator on many large and small subreddits also informed me that moderation becomes more demanding, complicated and involved as subreddits grow large. Therefore, for this study, I decided to focus on large subreddits, aiming to unpack how automated mechanisms help regulate large, active, long-running subreddits. To that end, I used a purposive sampling approach to select participants for this study (Merriam, 2002). The power of this sampling lies in selecting information-rich cases whose study illuminates the questions of interest (Patton, 1990). I used this approach to recruit participants who moderate large, high-traffic subreddits.

I began with a list of the 100 largest subreddits, as determined by their subscriber count, available on the RedditMetrics website [2]. I sampled subreddits from this list that are at least five years old, show high levels of activity (at least 100,000 comments posted in the period June 1 to Dec 31, 2017[3]) and reflect a diverse range of topics and moderation rules. My sampled subreddits are: r/photoshopbattles, r/space, r/explainlikeimfive, r/oddlysatisfying and r/politics (Table 5.1). I received permission from our participants to disclose the names of these subreddits. I chose not to anonymyze the names of these subreddits because knowing the identity of these subreddits is important to ground this research and help readers contextualize the nuances of my findings.

I observed about a hundred posts on each of these five subreddits and found that these communities reflect a wide range in the themes of conversations. These subreddits show

[2] http://redditmetrics.com/top

[3] I retrieved this data by running SQL-like database queries on the public archives of Reddit dataset hosted on the Google BigQuery platform (BigQuery, 2018).

some overlap in their submission guidelines (e.g., all five subreddits ask users to "be civil" or "be nice"). However, most guidelines on these subreddits directly reflect the focus and norms of the communities and are therefore quite unique to them. For example, r/photoshopbattles has a list of seven rules about the types of images that are allowed to be submitted, reflecting the focus of the subreddit on image submissions. To take another example, a majority of rules on r/explainlikeimfive focuses on the types of questions and explanations that are allowed to be posted on the subreddit. These subreddits also differ in how sensitive and emotionally charged the discussions are. Therefore, selecting these subreddits allowed me to gain insights into a diverse range of moderation practices that are related to the use of automated mechanisms.

5.2.2 Interviews

I interviewed three moderators from each of the five subreddits selected in my sample so that I could triangulate themes from multiple perspectives on the work of moderation for each subreddit and attain a deeper understanding. In addition, I interviewed Chad Birch, a Reddit moderator (past moderator of r/games; currently moderates r/automoderator) who created Automoderator. Iris Birman, my research partner in this project, assisted me in conducting these interviews. I invited moderators to participate in semi-structured interviews with me by contacting them through Reddit mail. I also met many Reddit moderators at CivilServant Community Research Summit, a gathering of moderators and researchers held at the MIT Media Lab in Boston in January 2018, and recruited some moderators present who moderated any of the five subreddits in my sample. Participation was voluntary and I did not offer any incentives for interviews.

In these interviews, to understand the role of Automod in context, I first asked participants more general questions, such as how they became moderators and what typical tasks they engage in while moderating. Next, I asked them questions about how they use automated moderation mechanisms on their subreddit. I inquired about the type of conversations they try to foster in their communities and how automated mechanisms help them attain their goals. I also discussed with our interviewees how they coordinate with other moderators to configure automated regulation tools and how the subreddit policies affect their use of these tools. Finally, I asked participants about the limitations of automated tools and the challenges they face in using them.

Each interview session lasted between 30 and 90 minutes, and was conducted over the phone, on Skype, or through chat. I contacted some participants for further clarification of their responses during the analysis stage. Although I attempted to get additional data such as Automod configuration rules and moderation log (described in Section 2.1.6), my participants were hesitant in providing me these data because they contain sensitive information. Even so, some moderators calculated and sent me the proportions of moderation work done by Automod in their subreddit (Table 5.3). Importantly, in their interviews with us, all our participants were forthright and provided us many specific examples of how they use Automod and how they enforce different subreddit rules. Therefore, I have primarily relied on the interview data to present my findings.

In addition to conducting interviews, I spent over 100 hours moderating multiple large (e.g., r/science) and small (e.g., r/youtubers) subreddits over the last year to understand the dynamics of Reddit moderation systematically. I supplemented my interview data with participant-observation field notes (Taylor, Bogdan, and DeVault, 2015) taken while moderating these subreddits.

5.2.3 Participants

Sixteen moderators participated in this study. All of my participants were male[4]. Fifteen participants reported being in their 20s or 30s, and one participant chose not to share his age. Although a majority of my participants are from the US, I also interviewed moderators from UK, Ireland, Canada, Australia, and India. Table 5.2 provides some demographic

[4] I discuss this limitation in Section 5.4.4.

and moderation experience related information about my participants. I use light disguise (Bruckman, 2006) in describing my participants in this table and in my findings. Therefore, although I have omitted sensitive details to protect the identity of my participants, some active members of their communities may be able to guess who is being discussed. I also note that I have not anonymized Chad Birch, the creator of Automod, after asking for his permission and to provide him credit for his work (Bruckman, Luther, and Fiesler, 2015).

I note that my approach to subreddit selection introduced some limitations in this study. Specifically, my results are based only on participants who moderate five large subreddits (in addition to the creator of Automod). Although my participants moderate a variety of large as well as small subreddits, my interviews mostly focused on moderation of large subreddits. Therefore, my findings should be interpreted as representative of moderation on large communities only. Even though I focused on only five Reddit communities, conducting independent interviews with three moderators from each community allowed me to check my participants' interpretations of events and processes against alternative explanations. It also helped me discover the properties and dimensional ranges of relevant concepts in my analysis (Straus and Corbin, 1998). I also stress that my participants were quite diverse in terms of their backgrounds, including their tenure as a moderator and the number and types of communities they moderate.

5.2.4 Analysis

I fully transcribed the data from the interviews and read it multiple times. Next, I applied interpretive qualitative analysis to all the interview transcripts and field notes (Merriam, 2002). This process entailed a rigorous categorization of data as I identified relevant patterns and grouped them into appropriate themes. My analysis began with "open coding" (Charmaz, 2006), in which I manually assigned short phrases as codes to my data. This first round of coding was done on a line-by-line basis so that codes stayed close to data. I gathered 481 first-level codes at this stage. Examples of first-level codes include "40-50%

Table 5.2: Study Participants. This table provides the following information about my participants: the subreddit in my sample that they moderate, their age, total number of subreddits that they currently moderate, and their tenure as a moderator on the corresponding subreddit. If participants did not want to share certain information or the data was unavailable, I have noted it with “NA”.

Subreddit	**Participant**	**Age**	**# of subs moderated**	**Tenure**
r/photoshopbattles	PB_1	33	91	2 years
r/photoshopbattles	PB_2	Late 20’s	5	4 years
r/photoshopbattles	PB_3	NA	7	5 years
r/space	$Space_1$	25	2	1 year
r/space	$Space_2$	20-25	NA	NA
r/space	$Space_3$	33	11	1 year
r/oddlysatisfying	OS_1	32	28	4 years
r/oddlysatisfying	OS_2	21	12	10 months
r/oddlysatisfying	OS_3	26	8	3 years
r/explainlikeimfive	$ELIF_1$	30	8	5 years
r/explainlikeimfive	$ELIF_2$	27	8	4 years
r/explainlikeimfive	$ELIF_3$	28	3	1 year
r/politics	Pol_1	32	8	3 years
r/politics	Pol_2	25	12	1 year
r/politics	Pol_3	25-29	5	1 year
r/Automoderator	Chad	34	8	6 years

of moderation action taken by Automod" and "Automod rule resulting in many mistaken removals."

Next, I conducted multiple subsequent rounds of coding and memo-writing. I engaged in the continual comparison of codes and their associated data with one another. My coauthors and I discussed the codes and emerging concepts throughout the analysis. After the first round of coding that closely followed the text, my next round of coding was more high level and resulted in codes such as "Refining Automod rules over time" and "Finding Automod to have a steep learning curve."

In subsequent rounds of coding, I combined and distilled my codes into seven key themes. These themes included "Use of automated moderation tools other than Automod" (discussed in Section 5.3.1), "Automod creation and incorporation into Reddit" (Section 5.3.2), "Utility of Automod" (Section 5.3.2), "Use of Automod to enforce community guidelines" (Section 5.3.3), "Social dynamics around the use of Automod" (Section 5.3.4), "Configuring Automod rules" (Section 5.3.4) and "Challenges of using Automod" (Section 5.3.5). In addition to the ones reported in this paper, themes such as "Becoming/continuing to be a moderator" and "Recruiting new moderators" emerged but were excluded in further analysis. Next, I present my findings.

5.3 Findings

I now present my findings from my interviews with Reddit moderators and my participant observations as a Reddit moderator. I begin with a discussion of how Reddit moderators build, use and share a variety of automated tools. Following this, I focus on the creation and use of Automod. First, I showcase the creation and incorporation of Automod into Reddit, and highlight the utility of Automod for moderation teams. Next, I outline the use of Automod for enforcing different types of community guidelines. I then present my findings on the mechanics of Automod configuration. Finally, I discuss how Automod creates new tasks for Reddit moderators.

5.3.1 Reddit Moderation Tools

Reddit moderators use a variety of moderation bots to automate removal of undesirable content. My participants told me that Reddit has an open and easy-to-use API that promotes the development of such bots. One popular bot checks for whether a submitter has 'flaired'[5] a post. Another popular bot called 'Botbust' identifies and bans Reddit bots that post spam or offensive content or comments that provide no value to the community ((Submitter), 2016). Yet another bot helps moderators use their phones for moderating tasks by looking out for specially formatted comment tags left by them.

> *"What they'll do for example is, they'll have a bot that looks for a moderator leaving a comment like, '!R1.' If it sees a comment like that made by a moderator, that means, remove this post for violating rule one… It lets the moderators do this big process of removing the post, leaving an explanation comment, everything like that that would be very tedious to do on mobile manually, just automatically by leaving a really short comment." - Chad*

I found that some moderators prefer to design and implement moderation bots from scratch. Five participants told me that certain moderators on their subreddits create bots themselves so that they can attend to the specific needs of the community. Pol_1 pointed out that many subreddits even recruit users who are adept at writing Reddit bots as moderators so that those users can help automate various content regulation tasks. In a similar vein, PB_2 noted:

> *"We have multiple bots that we have made ourselves to automate out tasks such as catch[ing] plagiarism and detect[ing] reposts."*

[5] Flairs are usually used to arrange submissions into different categories. For example, r/science is a popular subreddit where users post links to peer-reviewed research. Submitters are required to flair their posts by indicating the subject of the research, e.g., Chemistry, Astronomy, Paleontology, etc. Flairs allow readers to quickly filter for posts that they are interested in.

While some moderators build bots from scratch, others frequently use tools made by other subreddits. Reddit moderators often moderate multiple subreddits (see Table 5.2) and develop relationships with moderators of many communities. My participants told me that they use their connections with moderators from other subreddits to borrow tools and bots that improve content regulation for their own communities. For example, PB_2 said that r/photoshopbattles only allows submission of images with reasonable quality but the community did not have any tools to automatically check image attributes like image resolution. In order to automate checking image quality, the moderators of r/photoshopbattles borrowed a bot from another subreddit and deployed it on their subreddit. Similarly, OS_3 said:

> *"We are talking about using one of /r/technology bots that will help reduce the amount of spam that we get. This was someone's pet project so it has been tailored to specifically what we want."*

This sharing of moderation tools indicates that social interactions between moderators of different communities play a role in helping moderators discover and implement automated mechanisms for improving content regulation. Indeed, lack of such informal interactions may lead to duplication of efforts in creating moderation bots. Some participants told me that even though bots with similar functionality may have been developed by moderators of other subreddits, a lack of central repository of such bots forces them to develop their own tools from scratch.

> *"There is no binder with 'Oh, you want to do this? Here's the code for that!' So, you will often get duplicate efforts." – Pol_2*

A complementary set of regulation tools includes those that don't remove the content themselves but help moderators enact regulation tasks more efficiently. For example, a majority of my participants use Reddit toolbox [6], a browser add-on that provides many

[6] https://www.reddit.com/r/toolbox/

useful features such as allowing moderators to remove a comment and all its replies with a single click, and tagging pre-written reasons for removal when they remove a comment or post. Similar to many other Reddit moderation bots, Reddit toolbox is also a product of voluntary work of users dedicated to maintaining Reddit communities.

In summary, moderators on each subreddit use a wide variety of automated tools, largely developed by volunteer Redditors, to enact content regulation. Given that different communities have different moderation needs and require specific solutions, it is valuable for Reddit to have a third-party API that developers can use to build customized tools. Next, I turn to my findings on the introduction and use of Automoderator (or Automod). I dedicate the rest of my findings in this chapter to discussing this tool because it is the most widely used automated regulation tool on Reddit, and it immensely affects Reddit moderation.

5.3.2 Introduction of Automod on Reddit

All subreddits in my sample allow every post to pass through by default and only remove a post later if a moderator wishes to reject it. Although moderators have the capability to configure settings so that only posts that get approved by the moderators appear on the site, this arrangement is unsustainable on communities that experience heavy traffic.

> *"I don't know of any big subreddits that do it because it would become untenable to physically moderate every single post before it's seen." – Pol_2*

But allowing all posts to appear by default creates a potential for situations where undesirable content is not removed. Many subreddits have a limited number of human moderators who only moderate posts at certain times of the day. Before tools like Automod were available, subreddits often had posts that were offensive or that violated the community rules, but they remained on the site for hours despite many user complaints until a human moderator accessed Reddit and noticed them. This lack of moderation often agitated the regulars of subreddits. Thus, there was a need to create an automated tool that would remove at least the most egregious postings without human intervention.

Automod Creation and Incorporation into Reddit

The original version of Automod was voluntarily created by Chad Birch using Reddit API [7] in January 2012. Chad was inspired to create this tool when he was a moderator on the *r/gaming* subreddit. He noticed that many of the tasks he did as a moderator were mechanical, e.g., checking the domain name of submitted posts to see whether they belonged to any of the common suspicious sites, checking whether the post submitter was a known bad actor, and looking out for some keywords that indicated that the post should be removed. He felt that such tasks were amenable to be performed automatically. Following this, he built the original Automod as a bot that could be set up with conditional checks, apply these defined checks to all newly posted content and perform the configured actions such as post removal and user ban if the checks were met (Deimorz (Submitter), 2012). These checks could be defined using regular expressions, which allowed for defining patterns in addition to specific words. For example, one subreddit configured Automod using the following regular expression to catch and remove many homophobic slurs:

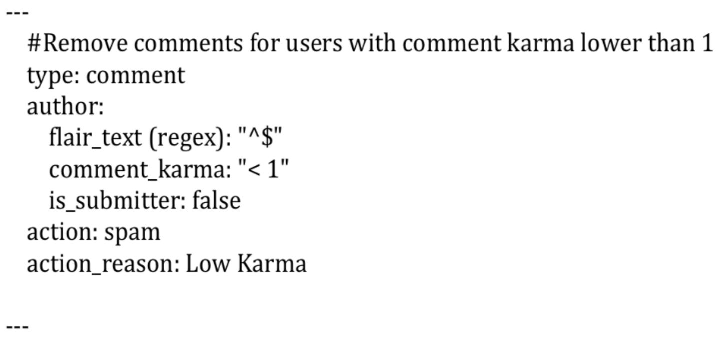
(ph|f)agg?s?([e0aio]ts?|oted|otry)

This single expression catches many slur words such as ‘phagot,’ ‘faggotry,’ ‘phaggoted,’ etc.

These check conditions could be combined in any manner and could also be inverted so that any content not satisfying the condition could be approved or removed. Using these basic building blocks, Automod could be used to develop a variety of capabilities (e.g., see Figure 5.1) such as banning posts from suspicious domains, auto-approving submissions from users whose account age and karma points [8] are higher than some threshold values, and removing user-reported comments containing certain phrases.

[7] https://www.reddit.com/dev/api/

[8] Reddit karma are digital reward points that users gain by posting popular content.



```
---
  #Remove comments for users with comment karma lower than 1
  type: comment
  author:
    flair_text (regex): "^$"
    comment_karma: "< 1"
    is_submitter: false
  action: spam
  action_reason: Low Karma

---
```

Figure 5.1: An Automod configuration snippet written in YAML format. This rule labels all comments posted by users having less than 1 comment karma score as spam and removes those comments, assigning 'Low Karma' as the removal reason. Exceptions include cases where the comment author is flaired (user flairs are digital markers usually assigned by moderators to trusted users) or where the user comments in a thread submitted by herself.

Chad found that a significant amount of moderation work could be handled using these basic configurations. He observed that implementing this tool in the r/gaming subreddit drastically reduced the amount of human moderation needed to regulate that subreddit. Seeing this, he offered the use of Automod to various other subreddits (Deimorz (Submitter), 2012). Thereafter, Automod quickly became popular on many communities.

Eventually, Automod was officially adopted by Reddit in March 2015 and offered to all the subreddits. Currently, each subreddit has its own wiki page for configuring Automod rules (Automoderator, 2018). This page is accessible only to the moderators of that subreddit. Moderators can define the rules for their subreddit on this page in YAML format (Ben-Kiki, Evans, and Ingerson, 2005), and these rules go into effect immediately after they are configured. Figure 5.1 shows an example of Automod rule that instantly removes comments posted by users with low karma. Reddit's official adoption of Automod further contributed to its popularity.

> *"When it [Automod] became an option on the site for the subreddit settings, it was infinitely easier to set up, and it became in-house, so it was a little bit more reliable." – Pol*$_2$

> *"I think, putting it in control of people directly made a big difference. . . they*

> *feel a lot better being able to know, 'Okay, I have this configuration. If it starts doing things that I don't want it to, I can just wipe this page out and it'll stop.' " – Chad*

Similarly, many other participants told me that they value the level of control that Automod provides them. Even when Automod mistakenly removes posts, they can usually recognize the syntactic setting that caused that removal, and change that setting to avoid similar mistakes in the future. In contrast, more advanced automated systems that use machine learning or neural nets may not be able to provide such specific causal understanding of their decisions to the moderators.

The popularity of Automod and its eventual official adoption by Reddit highlights the value of moderation tools developed by volunteers users. Next, I discuss how the introduction of Automod helped improve the efficiency of content regulation.

Utility of Automod

Automod can be configured to handle removal of undesirable submissions and comments separately. I discuss the automated removal of submissions and comments together in this chapter. This is done because my analysis suggests that they are both configured using similar regex patterns and they present similar benefits and challenges.

After Automod was introduced, moderators were able to enforce their subreddit rules more effectively and efficiently. For example, many moderators configured rules that auto-removed posts which received more than a certain threshold of user complaints. This allowed the moderators to regulate their subreddit even when human moderators were not proactively monitoring new posts and comments throughout the day.

> *"Reddit moves so quickly that once a post is a day old, it is just irrelevant to even moderate it at that point. It was really nice to have that automated in some ways... [*Automod*] allowed certain things like strict title formatting and stuff*

Table 5.3: Automod removal rate - % of removed comments (and submissions) that were removed by Automod over a month. These values were reported by my participants.

Subreddit	**For comments**	**For submissions**
r/oddlysatisfying	29.28%	4.95%
r/photoshopbattles	81%	66%
r/politics	79.86%	33.66%
r/explainlikeimfive	57.89%	72.78%

> *to become possible because before, a subreddit could never really have very strict title requirements or anything just because you would need a moderator to manually enforce those title requirements and it would never happen." – Chad*

All of my participants told me that they consider Automod an indispensable tool in their work, especially for moderating large subreddits. Table 5.3 shows the proportion of all removed submissions and comments that were removed by Automod on different subreddits over the period of a month, as reported by my participants. Although I couldn't obtain this breakdown between submissions and comments for the r/space subreddit, Space_3 told me that Automod does 40-50% of all moderation actions on the r/space subreddit. ELIF_1 pointed out:

> *"Extensive* Automod *rules is the only reason it's possible to moderate ELIF with the number of people [human moderators] we have been able to get to help us."*

Similarly, Pol_2 said:

> *"It's so powerful! I mean, most subreddits have thousands of lines of this code that takes a lot of the menial work out of it ... If it was to go away at some point, subreddits would become horribly moderated, and basic things would just grind to a halt."*

Some participants noted that since many inappropriate postings on Reddit are made by new users who may not be aware of the rules and norms of the community, a key advantage of Automod is that it can be used to gently nudge these users in the correct direction and to influence them to confirm to the standards of the community. This is because moderators can configure Automod to provide explanations for content removals to the users who posted them. Figure 5.2 shows an example of an explanation comment posted by Automod. My participants told me that such explanations usually provide detailed descriptions of the subreddit rule the submitter has violated and the steps that can be taken to avoid such removals in the future. Participants noted that such explanations are often effective in helping users understand the social norms of the community. For example, PB_2 said:

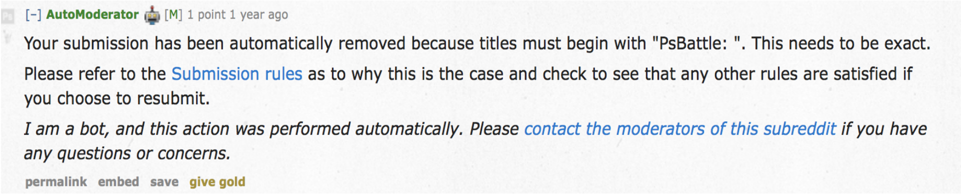
> *"A lot of the time, it is just easier to follow the rules, and people that don't want to waste their time tend to conform to my requirements rather than go through the hassle."*

I will elaborate on this point again in Chapters 6 and 7, when I discuss the use of automated tools to provide removal explanations.

Over time, Automod has become an integral and indispensable part of the content regulation system on Reddit. It executes a large amount of menial work previously done by human moderators, and helps new users understand the norms of the subreddit. Next, I discuss how the syntactic configurations of Automod allow subreddits to automatically enforce some of their submission guidelines but not others.

5.3.3 Use of Automod to Enforce Community Guidelines

Community guidelines play an essential role in shaping the community and its moderation. They not only establish standards for how the users should behave on the subreddit, but they also set expectations for what users can count on from the moderation team. These guidelines (often called subreddit rules) directly affect how Automod is configured. That



[–] AutoModerator [M] 1 point 1 year ago

Your submission has been automatically removed because titles must begin with "PsBattle: ". This needs to be exact.

Please refer to the Submission rules as to why this is the case and check to see that any other rules are satisfied if you choose to resubmit.

I am a bot, and this action was performed automatically. Please contact the moderators of this subreddit if you have any questions or concerns.

Figure 5.2: A comment posted by Automod in response to its removal of a submission on r/photoshopbattles. Automod on r/photoshopbattles has been configured to automatically detect submissions that do not begin with ‘PsBattle:’, remove those submissions, and post this message as reason for removal.

is, Automod rules are often created so that they can help ensure compliance with these guidelines.

> *“We have our rules in place and we make Automod conform to them as much as possible.” – ELIF$_2$*

For example, the submission guideline #1 on r/photoshopbattles says, “All titles must begin with PsBattle:” [9] To ensure that users comply with this rule, moderators on r/photoshopbattles have configured an Automod rule that detects and removes all submissions that don’t begin with ‘PsBattle:’ (see Figure 5.2). Here, it is worth noting that Automod not only removes an undesirable submission, it also automatically provides the poster the reason for removal, thereby serving an educational role (West, 2018). Similarly, Automod helps ensure compliance with the r/explainlikeimfive guideline #10: “All Posts Must Begin With “ELI5.”

> *“We already were reasonably strict in our rules, and a few of them were very easy to implement as Automod rules right out of the gate. The most obvious is that posts must start with “ELIF:” in our sub. We frequently removed posts without that prefix and asked them to repost it, and then suddenly Automod could do that perfectly and instantly with no work.” – ELIF$_1$*

As another example, the r/politics guideline # 3: “Submissions must be from domains on the whitelist” requires that users should only post links from a specified list of web-

[9] This guideline helps distinguish r/photoshopbattles images from other pictures when they appear on the Reddit front page along with images from other subreddits.

sites that are considered appropriate for the subreddit[10]. The r/politics moderators ensure compliance with this guideline by configuring an Automod rule that checks whether each submission links to one of the domains on the whitelist. Another r/politics Automod rule guarantees compliance with the submission guideline #9: "Do not use "BREAKING" or ALL CAPS in titles" by checking for each submission whether its title contains all uppercase letters or if it contains the word "BREAKING."

While the above examples illustrate cases where Automod rules were created to enforce compliance with the existing guidelines, I also found a few instances where subreddit guidelines were specifically created to make content regulation easier to operationalize using Automod. For example, r/oddlysatisfying created a rule that requires submitters to describe the content shown in the image they submit in their post title. Enforcing this rule provides more information for Automod to detect and remove posts that are unsuitable for the r/oddlysatisfying community. Participant OS_1 explained how this rule helped ease the workload of moderators:

> *"The title rule helped make it easier to mod[erate] as it sped up the process of removing something that was NSFW[11] or had no purpose in the sub without having to look at or click every single submission."*

Here, we see how a community guideline was created to facilitate automated processing and alleviate the work of content moderators. This guideline, however, increases the input required of end-users by requiring them to provide a description of their image submissions. This is an example of how adopting automated tools to ease the work of moderators can create additional burdens on other stakeholders.

I found some variations in how different subreddits configure and use Automod. In Section 5.3.2, I showed the differences in the proportion of actions taken by Automod on various subreddits. My interviews provided additional insights on this point. For example,

[10] This whitelist is available at https://www.reddit.com/r/politics/wiki/whitelist.

[11] Not Safe For Work

OS_2 told me that r/oddlysatisying uses Automod only to detect "bad and threatening language." In contrast, PB_1 informed me that r/photoshopbattles configures Automod to help facilitate more sophisticated content curation, e.g., "Photoshops Only Mode" threads that only allow comments that are photoshopped versions of the image in original submission. In this case, Automod is configured to remove any comments that don't link to an image hosting site.

> *"I set the thread to "Photoshops Only Mode" when the thread reaches 150 comments. That way, AutoModerator takes care of off-topic chains for me." – PB_1*

Not all policy guidelines are equally amenable to be enforced using Automod configurations. Moderators of every subreddit in my sample pointed out the grey areas in their guidelines that require subjective interpretations, and they consider such guidelines harder to implement using Automod. For example, $ELIF_3$ told me that r/explainlikeimfive moderators do not rely on Automod to enforce four (out of ten) subreddit guidelines[12] because they require a level of interpretation that are beyond the capabilities of Automod, e.g. guideline #4: "Explain for Laypeople" and guideline #5: "Explanations Must Be Objective." Instead, human moderators review all comments to ensure that these guidelines are followed.

A majority of my participants also noted that Automod rules are unable to consider context when making moderation decisions. For example, certain words or phrases can be used in multiple settings – their use may be acceptable to moderators in some contexts but not in others. $Space_1$ described how Automod removed a comment containing the term "shit" that he had to manually approve:

> *"Somebody talked about how they've "read this shit" in an explanation of their longstanding fascination with a topic. People that use that word tend to use it in mean-spirited or unserious comments, but this was an example where it's just for emphasis." – $Space_1$*

[12] https://www.reddit.com/r/explainlikeimfive/wiki/detailed_rules

A few participants told me that their subreddits prohibit hate speech, but when moderating content posted to the subreddit, they sometimes find it difficult to determine whether a given comment should be considered hate speech or not. They try to attain a balance between allowing users to freely exchange ideas and prohibiting dialogue that make some users feel attacked or unsafe. In pursuit of this balance, however, participants find it challenging to use Automod to enforce guidelines prohibiting offensive behavior. For example, ELIF$_1$ shared:

> *"Our #1 rule is "Be nice," and we take that very seriously. That's definitely something that requires some interpretation. Someone saying "Way to be an idiot" is not nice. But is someone saying, "the way this works is X, people who think it works like Y are idiotic" not nice?"*

Knowing these limitations, moderators generally tend to configure Automod in a way that allows decisions that are less ethically ambiguous to be made automatically. Posts that are caught by Automod are removed as soon as they are posted, but many undesirable posts remain that are later removed by human moderators when they are reviewed. As a result, the use of Automod offloads some of the work of human moderators and frees them up for other tasks, e.g., making moderation decisions on posts that are harder to adjudicate.

> *"The goal of Automoderator is to get rid of the clear-cut things that you don't need a human to do. You can just look at the text and go, hey yeah, it's against the rules." – Pol$_2$*

> *"Automod does a lot of filtering of the worst stuff for us...It makes things easier and less stressful. We don't have to be trolling every thread for the worst stuff to get removed." – Space$_1$*

As the above quotes indicate, the use of Automod not just decreases the total amount of work that moderators need to do, it also reduces the emotional labor of moderators by minimizing their exposure to violent or offensive content.

Automod can be configured to either directly remove a post or triage it to a queue where human moderators can review that post. When Automod is used to flag a post for human review, the default decision (configurable in Automod) can be to either allow the concerned post or to automatically remove it until a human moderator gets to review that content. Therefore, the use of Automod provides some flexibility even when making difficult decisions. For example, OS_2 described a scenario in which Automod flags accounts for human review:

> *"Basically if a user's account is under X days old or has less than X karma, it will be automatically placed in the report queue to make sure it's not a spammer." – OS_2*

To sum up, Automod is well adept at enforcing some of the policy guidelines but not others. Subreddits still have to depend on human moderators to enforce guidelines that require subjective interpretation. This highlights the gap between what the syntactic instantiation of Automod rules can do and what the subreddit policies that are at a semantic level require. Next, we discuss the factors that influence how Automod is configured.

5.3.4 Mechanics of Automod Configuration

Given that adding or changing any rule of Automod can affect the moderation status of a large number of posts, I explored in my analysis how the decisions to edit Automod rules are made and who makes these edits.

Social Dynamics

My interviews suggest that only a few moderators in each subreddit take on the responsibility of actively configuring Automod rules because it is difficult for others to understand how to configure it.

> *"All the active mods in ELIF can edit Automod, but few do because it's complicated, and our Automod config is pretty huge." – $ELIF_1$*

Participant PB_2 told me that he has handled most of the Automod coding of r/photoshopbattles for the past 4.5 years. Although that subreddit has a couple of other moderators who know how to program Automod, the rest of the moderation team considers PB_2 a single point of control for managing Automod code. PB_2 described that this arrangement allows the subreddit to identify and debug errors in Automod configurations more efficiently. He elaborated:

> *"I am a single point of control because [otherwise] if Automod starts going wrong then I have no idea what the issue is or how to fix it . . . Standards and good general practice make things easier to maintain but any time you get a group of people working on the same code without strict control, things get messy. It has just happened that I just handle it."*

In line with this, some participants from other subreddits told me that they are not informed of most of the changes in Automod configuration. They stressed, however, that they have the authority to reverse the decisions made by Automod if they deem them to be in error. For example, $Space_1$ stated:

> *"I am not always consulted as to what Automod will filter, but I can always override Automod."*

Six of my participants reported editing Automod rules. Participants with little knowledge or prior experience of editing Automod reported making only minor changes to existing Automod rules. For example, if there is an Automod rule that removes posts containing any word from a list of swear words, novice moderators feel comfortable adding another swear word to that list. This is because, as my participants pointed out, they have noticed Automod removing comments containing any swear word already on the list. They therefore recognize that their addition of a new swear word to the Automod rule will activate similar removals of all subsequent user comments containing that word. These moderators restrict themselves to making only minor changes, however, so as to ensure that their changes to Automod do not create errors in the functioning of Automod or result in unanticipated moderation actions. For example, OS_3 said:

> *"I don't get too involved with Automod. I know how to do the basic stuff, but not the coding side of things."*

In addition to restricting themselves to making only minor changes, novice moderators also usually inform the entire moderator team of the changes they make to Automod rules so that moderators with more experience of editing Automod can undo those changes, if required. Participants also told me that when they make changes to Automod rules, they often add documentation on the reasons for those changes as comments in the Automod code. Thus, moderators self-calibrate their skills at configuring Automod and practice care when making any changes to Automod rules.

As another example of moderators practicing care in their work, moderators often decide how to add or change any Automod rule based on their expectations of the number of posts that rule will affect. For example, Pol_1 told me that when he makes changes that are not expected to affect too many additional posts, he makes such configuration edits either by himself or by consulting with a few moderators. On the other hand, when he adds another domain to the white list of domains[13], this addition is discussed and voted upon by all the moderators before it is configured in Automod. This is because it is expected that many new posts may link to the domain in question and adding a rule to allow such posts may substantially affect the content available on the subreddit.

Participants from each subreddit in my sample noted that the moderator teams sometimes deliberate over the issues around the configuration of Automod rules in communication channels like Slack and Discord. However, such discussions are relatively rare. For example, $ELIF_1$ said:

> *"I don't think we've really had a discussion about Automod in a while. I see*

[13] As mentioned before, r/politics subreddit has set up a rule in Automod to allow only those submissions that link to one of the configured whitelist of domains.

> *it like a mop for janitors. It's necessary, and you use it all the time, but there isn't really much to discuss about it."*

Half of my participants complained that Automod has a significant learning curve. Even moderators who understand regular expressions do not often use the advanced capabilities of Automod such as checking multiple conditions before triggering a rule. They either do not realize such configurations are possible or consider crafting such configurations too difficult. For example, Pol_1 said:

> *"It's not necessarily user-friendly. . . it almost entirely functions on regex, and its own little quirks and syntax to implement things so it can take some time for people to get decent at using it. There are a lot of mods whose eyes glaze over when having to work with it and [they] would rather do something else."*

As a result of the steep learning curve of configuring Automod, some moderators with little or no prior experience of editing Automod rely on other moderators for making the requisite changes. For example, $ELIF_2$ told me that moderators who do not know how to configure Automod often request other expert moderators to make changes in the configuration. $ELIF_2$ argued that complying with such requests added new responsibilities over moderators who take on the job of editing Automod without conferring any additional power or benefits to them. He further explained:

> *"Honestly, I feel like the mods who know more about bots get run over a bit. "Hey we need to add this, will @xyz or @yzz help please?" on Slack, or whatever."*

I found that moderators do not reveal the details of exactly how Automod works to their users. For example, the wiki page for Automod rules is not accessible to regular users by default. My participants told me that although Reddit provides them the ability to make this wiki page public, they choose not to do so to avoid additional work and to ensure that bad actors do not game the Automod rules and post undesirable content that Automod cannot detect.

> *"If you know what exactly is in the [Automod] code, then it is easily bypassed/exploited. Users with the inclination could figure out how to attack us maliciously or spam unhindered." – PB$_2$*

> *"It would be a massive pain in the ass to have it public, because it would require regular published updates, and we'd end up having to explain each modification to at least one of the few people complaining about censorship." – ELIF$_2$*

This necessary lack of transparency creates tensions between the moderators and the community. Many users are not aware of the presence of Automod, and posters whose comments are removed are usually not shown whether their comment was removed by Automod or a human moderator. Therefore, when users' comments get mistakenly removed by Automod, they often attribute such removals to human moderators and consider them unreasonable. Participant Pol$_2$ explained:

> *"It will be this big, incredible comment, and 98% of it will be, boy, this guy did his research, and he's doing good work, but it will get caught [by Automod] by something he says, and get removed. In those cases, it can turn into users thinking, 'Some moderators are filtering my speech — they don't like what I'm saying!' "*

By contrast, I observed that on many subreddits, when Automod removes a submission, the poster is notified that their submission was removed automatically. This is usually done through a comment to the removed submission authored by Automod (e.g., see Figure 5.2) or through automatically flairing the submission with a short removal reason (see Chapter 7). Thus, moderators negotiate in distinct ways which aspects of the use of Automod they reveal to their users and how. This is an example of added responsibility and decision-making that moderators are obliged to perform when they adopt Automod.

Need for Careful Curation of Automod Rules

A majority of my participants felt that the utility of Automod largely depends on how its configuration rules are set up. Some moderators who are not too familiar with using regular expressions end up writing rules that are too broad and remove content that they did not intend to remove.

> *"Sometimes, mods implement rules that accidentally remove too many things. In those cases, after a user has asked us to review a removal, I've gone back and refined the [Automod] rule to better capture what we're looking for. Regex (what the rules are made with) can be tricky." – ELIF$_1$*

Automod does not provide any feedback on the number of times a specific rule has been triggered. If the moderators do not pay attention to how their Automod rules affect the moderation on the subreddit by tracking the content that is being automatically removed, it may take them a while to realize the occurrence of unintended post removals. Seven participants from four different subreddits expressed their desire to get additional information about how Automod operates. They argued that analyzing statistical data about how different rules of Automod affect content regulation would allow them to fine-tune Automod configurations more efficiently.

> *"A poorly phrased regex bit can make something that looks like it shouldn't trigger on a post, trigger. But ... how do I know which one of the thirty five Automod rules did it? How do I know which part of the post made the trigger? ... I want to know which rules were invoked for which posts, how frequently, etc. - both in aggregate and on individual posts." – ELIF$_2$*

> *"If there was an easier way to see each removal reason and just a sampling of the comments that were removed for that removal reason, that would be pretty powerful. It would give you a better way to check your work." – Pol$_2$*

Moderators often rely on user reports to understand when an Automod configuration triggers too many unintentional side-effects. Although such user reports can allow moderators to correct their configurations, such mistakes create dissatisfaction among users.

> *"There was also a lot of users that were quite upset about it simply because they call it basically the censorship bot because it can just remove anything immediately with no ability for people to reason with it or convince them that it's the wrong decision." - Chad*

Three of my participants showed concern that many moderators focus on minimizing false negatives rather than false positives. In other words, moderators try to ensure that Automod is configured to catch as many undesirable posts and comments as possible, but they do not pay enough attention to whether Automod removes content that should not have been removed.

> *"I think, one of the things that bothered me before I was a moderator and complained about a lot was the false positives. Like, half the time, the Automoderator would have automatic comment removals if your comment had the word "homo" in it. But homo is not just a gay slur; it's also the genus of human beings – homo sapiens – so if you would write a comment using the word "homo sapiens" in it, your comment would be removed." – Space$_3$*

Here, we see that the moderators focus more on bad actors being punished, but ignore cases where good members are wronged. But the health of a community may be reliant on the latter just as much as on the former because undeserved punishment risks creating chilling effect and ultimately drives members away (as we saw in Chapter 4). Additionally, as we will discuss below (Section 5.3.5), false positives also result in many user complaints, which consequently increases the moderators' workload of responding to those complaints.

In summary, since Automod only allows configuring keyword-based rules, the moderators have to make tradeoffs between (1) removing all posts that use the keyword in question including the posts that are acceptable or (2) not constituting a rule for that keyword and manually removing posts that are unacceptable, thereby increasing the work of human moderators. Thus, the use of Automod complicates the value-laden issues involved in content regulation.

5.3.5 Automod Creates New Tasks for Moderators

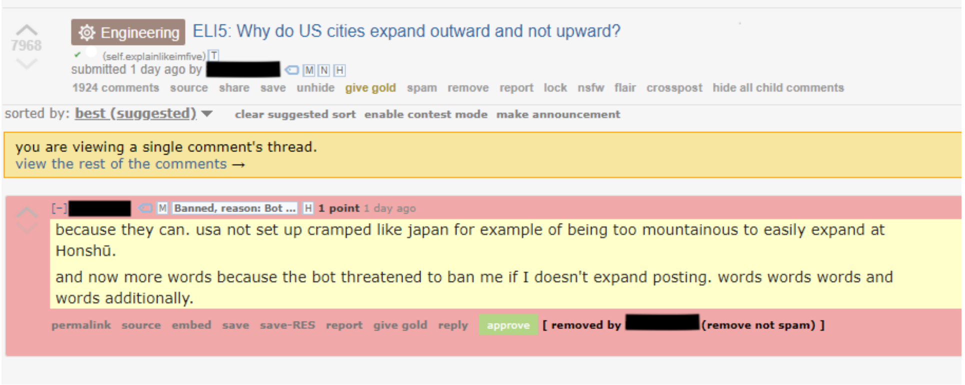
Although Automod certainly helps moderators to deal with the challenges of content regulation, its efficient deployment requires the moderator team to take on a set of new tasks in addition to configuration of Automod rules. I discuss these tasks in this section.

Regular Updating of Automod Rules

Participants told me that Automod rules on their subreddit have become more refined over time as the moderators continue tuning them to attain more accurate automated regulation decisions. Still, updating these rules is a continuous process - user content changes with the influx of new users and changes in cultural trends, and new requirements for automating specific moderation tasks are identified and configured for in Automod. Thus, Automod incorporates the historical evolution of the expectations for postings on each subreddit.

> *"95% of our Automod [code] changes are probably just adding, removing or refining items based on current events." – ELIF$_1$*

> *"Modifications to Automod code tend to come up on an as-need basis. If something slips by an existing rule, then the code for it is bolstered to cover that hole." – PB$_2$*


7968
Engineering
1924 comments source share save unhide give gold spam remove report lock nsfw flair crosspost hide all child comments
sorted by: best (suggested)
clear suggested sort enable contest mode make announcement
[–]
M
Banned, reason: Bot ...
H
permalink source embed save save-RES report give gold reply
approve

ELI5: Why do US cities expand outward and not upward?

(self.explainlikeimfive)

submitted 1 day ago by [redacted]

you are viewing a single comment's thread.
view the rest of the comments →

[redacted] 1 point 1 day ago

because they can. usa not set up cramped like japan for example of being too mountainous to easily expand at Honshū.

and now more words because the bot threatened to ban me if I doesn't expand posting. words words words and words additionally.

[removed by [redacted] (remove not spam)]

Figure 5.3: This comment shows the commenter's attempt to avoid detection by Automod. When ELIF$_1$ noticed this comment, he banned the commenter from r/explainlikeimfive subreddit. I have blotted usernames in the figure to preserve anonymity.

The wiki page for Automod rules records the history of each edit. This allows each subreddit to keep track of which moderators make what modifications to its Automod configuration. It also helps ensure that moderators who make any changes to Automod rules are accountable for their actions. It additionally allows quick reversion of any Automod rule changes, if needed.

Preventing Users from Circumventing Automod

My participants recognized that Reddit users can easily evade Automod rules by identifying the rule that triggers removals and use tactics like creative misspellings to bypass that rule. Therefore, they make efforts to guard against deliberate circumvention of Automod rules. As I discussed in Section 5.3.4, moderators do not provide users access to the wiki page where Automod rules are configured. Moderators of r/explainlikeimfive and r/photoshopbattles told me that they ban users who try to evade being caught by Automod. For example, ELIF$_1$ banned the user who posted the comment shown in Figure 5.3 because that user clearly attempted to bypass the Automod rule for removing any comments that are too short in length. In fact, some Automod rules on r/explainlikeimfive are configured to detect attempts to evade other rules and notify the moderators so that they can take appropriate actions against the suspected user.

Correcting False Positives

As I discussed in Section 5.3.4, moderators tend to focus on minimizing false negatives rather than false positives. As a result, Automod can often remove postings that do not violate community guidelines. My participants reported that a bulk of their moderation mail contained complaints from new users about mistakes made by Automod. This creates additional work for the moderators by requiring them to respond to user complaints about the false positives of Automod's decisions.

> *"Valid complaints are usually about our bots getting a false positive for their issue as we have them set up pretty tight to make sure certain things don't slip by." – PB_2*

Thus, while using Automod allows moderators to offload a lot of their work and enact content regulation more efficiently, it also requires moderators to develop new skills like configuring Automod rules and conduct additional activities like defending against deliberate avoidance of Automod filters and correcting false positives. Since these new tasks are not trivial, the use of Automod creates new challenges of training and coordination among moderators.

In summary, although Reddit moderators can regulate their communities without using any automated tools, the use of these tools makes their work more convenient and manageable. The combination of natural human abilities of moderators with the capacities of external components like Automod and Reddit toolbox forms a system that performs the existing function of content regulation more efficiently (Kaptelinin, 1996). In their research on the use of automated tools in Wikipedia governance, Geiger and Ribes noted that "the delegation of certain tasks to these tools makes certain pathways of action easier for vandal fighters and others harder" (Geiger and Ribes, 2010). Similarly, I found that using automated tools on Reddit changes the underlying activity of content regulation and raises new challenges (such as preventing users from evading Automod rules) that moderators need to grapple with.

5.4 Discussion

This research extends prior work on content moderation by drawing attention to the automated regulation tools that moderators use. I describe the sociotechnical practices that shape the use of these tools. I also highlight how these tools help workers maintain their communities. My analytic approach has allowed me to identify the limitations of current automated systems and recognize the important design challenges that exist in attaining successful moderation. In this section, I describe these challenges and limitations. I also propose solutions that may help address them. Furthermore, I discuss what other communities that incorporate automated tools in their regulation systems may learn from this research.

5.4.1 Facilitating Development and Sharing of Automated Regulation Tools

Centralized moderation tools and mechanisms are often developed using universalist design principles and practices that assume that the 'default' imagined users belong to the dominant social groups (Costanza-Chock, 2018). Yet, these official moderation tools may not be able to satisfy the requirements of all communities (Chapter 4). Moderators are well-poised to identify these social-technical gaps (Ackerman, 2000) because they work closely with their communities and can recognize the specific needs of their users that official moderation tools do not satisfy. Therefore, mechanisms that allow these moderators to develop and deploy regulation tools that meet the unique requirements of their communities can substantially improve content regulation.

One such mechanism is to provide moderators an API access to the community data. My findings suggest that the open and flexible API provided by Reddit platform has encouraged the development of a wide variety of automated tools. Volunteer users frequently create moderation bots that tailor to their community and improve its regulation. This is in line with my observation in Chapter 4 that Twitter blocklists, a third-party moderation tool developed by volunteer users on Twitter, helped enhance the experiences of many marginalized users by allowing them to curate content in ways that were not possible through centralized moderation mechanisms officially offered by Twitter. Therefore, I recommend that more platforms should consider providing API access that volunteer developers can use to build and deploy automated regulation bots that meet the specific needs of their communities. Opening up content regulation to third-party developers should also encourage the implementation and testing of creative new ideas for community management.

I found that Reddit moderators spend considerable time and efforts developing bots to improve content regulation for their own subreddit using automated mechanisms. But bots that are valuable for one community can also bring immense value to regulation of other communities. For example, I saw that although Automod was initially developed for the r/gaming subreddit, it eventually became an indispensable part of the Reddit regulation system. Still, as I discussed in my findings, there is no central repository of all the automated tools that moderators can directly use. Moderators only come to know about such tools through their contacts with moderators in other subreddits. This results in duplicate effort on the part of bot developers. To avoid such duplication, platforms like Reddit should encourage volunteer developers to build tools that can be quickly adapted to enact regulation in other similar settings. Platforms may also promote sharing of such tools on a centralized repository so that other moderators can directly access them and adapt them for their own communities.

5.4.2 The Need for Performance Data

My findings show that there is lack of accessible data on how well the automated parts of regulation systems on Reddit work. Currently, Automod does not provide any visibility into the number of times each rule has been triggered. This is problematic because rules added to Automod sometimes have unintended effects. I found that because of the absence of easy visibility into Automod behavior, moderators often have to rely on user reports to identify the occurrence of unintended post removals. This highlights the importance of flagging mechanisms (Crawford and Gillespie, 2016) and contributions of users to content regulation (Lampe and Resnick, 2004) even in systems that rely on automated tools. Yet, as my findings show, mistakes made by Automod can frustrate the users affected.

To facilitate quick discovery of Automod mistakes, I recommend that designers build audit tools that provide moderators visibility into the history of how each Automod rule affects the moderation on the subreddit. Such visibility would allow moderators to edit Automod rules if required and control its actions more closely. Audit tools could also be enhanced to show moderators the potential consequences of creating a new rule by simulating application of that rule on already existing data in sandpit type environments. If moderators are able to visualize the type of comments that would be removed by creation of a new rule, they would be better positioned to avoid crafting broad rules that result in many false positives.

I expect that the same principle also applies to automated regulation on systems other than Reddit. It is vital for human moderators to understand how different configurations of automated regulation systems affect the curation of their sites. As my findings show, moderators want the ability to quickly locate the settings that result in undesirable regulation decisions and fix them. Therefore, automated systems should be designed so that moderators have detailed visibility into how automation affects content curation. Moderators should also be able to tune the configurations of such systems at a granular level and maintain control over how these systems work.

While the provision of performance data, as discussed above, is important to evaluate automated moderation, it should be noted that systematic evaluation of moderation records, in general, is a non-trivial endeavor. A thorough evaluation would require sampling and reviewing of posts that have been allowed to be kept up as well as posts that have been removed - either by automated tools or by a human moderator. For each moderation action, different moderators may have different views of whether that action is appropriate - such conflicts and disagreements are part of the political process of enacting moderation. In addition, it is possible that a majority of users may disagree with the moderators' decisions. Thus, post-hoc evaluation of content moderation records as a social practice has many critical concerns that need careful examination, and is a ripe area for future research.

I also found that only a few technically adept moderators can configure Automod, and many subreddits are unable to tap into the full potential of Automod. This is similar to Geiger and Ribes' finding on automated regulation in Wikipedia that while many "workarounds are possible, they require a greater effort and a certain technical savvy on the part of their users" (Geiger and Ribes, 2010). Therefore, I recommend that automated systems be designed in such a way that moderators can easily understand and configure their settings. This would allow more moderators to engage with automated systems, and facilitate conditions where a larger share of moderators can influence content curation using automated tools.

5.4.3 Human versus Automated Moderation Systems

My analysis shows that identifying tasks that should be automated and configuring tools to perform those tasks is crucial for Reddit moderators' ability to maintain their communities, especially as the communities grow large. This finding is consistent with Epstein and Leshed's observation that automation of maintenance tasks like detecting and addressing incivility have been critical to scaling up the RegulationRoom[14] deliberation environment (Epstein and Leshed, 2016). While automation is crucial to support the growth of online communities, accurate automated detection is not an easy task. It is arguably impossible to make perfect automated moderation systems because their judgments need to account for the context, complexity of language and emerging forms of obscenity and harassment, and they exist in adversarial settings (McDaniel, Papernot, and Celik, 2016) where they are

[14] http://regulationroom.org

vulnerable to exploitation by bad actors.

Automating moderation not only facilitates scalability, it also enables consistency in moderation decisions. In a recent Data & Society report, Robyn Caplan noted that many companies tend to formalize their logic in order to address regulation concerns more consistently (Caplan, 2018). These companies transform content moderation standards into hard-coded training materials for new workers as well as automated flagging systems. This is in line with how moderation criteria on Reddit are specified as fixed regular expressions in Automod rules. Although hard-coding moderation criteria facilitates scalability and consistency of moderation systems, such transformation of content moderation values can end up being insensitive to the individual differences of content, for example, when distinguishing hate speech from newsworthiness (Caplan, 2018). These failures to address context issues can have serious consequences, e.g., persistence of misinformation campaigns on Facebook or WhatsApp that arguably contributed to violence in Myanmar (Stecklow, 2018).

More advanced automated systems that use machine learning models, especially those based on deep-learning frameworks, currently cannot provide specific reasons for why each removed comment or post was removed. Advances in human-understandable machine learning may help address this problem in the future (Lakkaraju, Bach, and Leskovec, 2016). Currently, my findings show that moderators adopt Automod because they can directly control how it works by editing its configuration. They can understand the mistakes made by Automod by observing the keywords that triggered those mistakes and explain such mistakes to placate dissatisfied users. Prior research has also shown how retaining control over content regulation is important to the moderators (Diakopoulos and Naaman, 2011).

Therefore, I caution researchers and designers that although AI moderation systems are invaluable for managing many moderation tasks and reducing workload (Soni and Singh, 2018), deploying such systems without keeping any humans in the loop may disrupt the transparency and fairness in content moderation that so many users and moderators value. This is in line with speculations made by other researchers that machine-learning driven moderation approaches are inherently risky because they may “drive users away because of unclear or inconsistent standards for appropriate behavior” (Seering, Kraut, and Dabbish, 2017). Additionally, I found that only a small number of moderators in each subreddit configure Automod because others do not have the technical expertise to make such configurations. The use of more complex machine learning tools can further raise the bar for users who can moderate online communities while also disproportionately increasing the workload of moderators who can work with those tools. Thus, platforms should consider that they may lose valuable moderators when moving to systems that heavily rely on machine learning tools.

Designers and moderators must recognize that the use of automated regulation systems fundamentally changes the work of moderators. For example, when subreddits use Automod, moderators’ work becomes constrained to adjudicate only those postings that are not caught and removed by Automod. Moreover, it creates additional tasks that require technical expertise such as regular updating of Automod rules and preventing users from circumventing Automod. I confirm the finding of Seering et al. (Seering et al., 2019b) that Automod sometimes adds to the work of moderators because they have to manually approve content mistakenly removed by Automod. Therefore, when moderators incorporate new automated mechanisms in their content regulation systems, they should anticipate new tasks and prepare to execute and train for those tasks. More generally, while moderators and new automated systems may co-evolve and adapt to each other, it is still important to consider the resulting social-technical gaps as CSCW problems and make efforts to “round off the edges” of coevolution (Ackerman, 2000; Postman, 1992).

I found that Reddit moderators show some aspects of the work of Automod to their users but not others. These decisions are important in order to retain the trust of the users while at the same time ensuring that bad actors do not game the system and bypass Automod rules. Concerns about users evading automated moderation systems are not limited to Reddit — prior research has shown how bad actors on Instagram and Tumblr circumvent platforms' efforts to moderate problematic hashtags by devising innovative ways to promulgate controversial content (Chancellor et al., 2016; Gerrard, 2018). Therefore, when incorporating automated regulation systems, moderators should be prepared to make critical decisions about which parts of the automated system to show and which to hide from their users.

My findings also bring attention to the tradeoffs between reducing the work of human moderators and not automatically removing posts that may potentially be valuable despite having suspicious characteristics. On one hand, using Automod reduces the amount of work that Reddit moderators need to do and protects them from the emotional labor of scrolling through the worst of the internet's garbage. On the other hand, it is all too easy for moderators to configure rules that are too broad in Automod. Although such a configuration catches and removes many potentially unacceptable posts and reduces the dependency on human moderators, it results in many false positives that may alienate users. I also found that human moderators are needed to frequently update Automod rules so that Automod can account for the fluidity of culture and adaptability of violators seeking to avoid detection.

Improving Mixed-Initiative Systems

Given the deficiencies of automated tools and the importance of careful human administering of these tools, I propose that instead of developing fully automated systems, researchers and designers should make efforts to improve the current state of mixed-initiative regulation systems where humans work alongside automated systems. Since automated tools are likely to perform worse than humans on difficult cases where understanding the nuances and context is crucial, perhaps the most significant consideration is determining when automated tools should remove potentially unacceptable material by themselves and when they should flag it to be reviewed by human moderators. It is critical for these tools to attain this balance to ensure that unintended post removals are avoided and at the same time, the workload of human moderators is substantially reduced. I echo calls by previous studies for building systems that ensure that the interactions between automation and human activities foster robust communities that function well at scale (Geiger and Halfaker, 2013; Epstein and Leshed, 2016; Seering et al., 2019b).

Another promising direction to explore would be to build systems that adapt hybrid crowd-machine learning classifiers like Flock (developed by Cheng and Bernstein (Cheng and Bernstein, 2015)) for the purpose of content regulation. Such systems would require a dataset of comments that have been thoroughly moderated and labeled as ‘approved’ or ‘removed’. To begin with, such a system would guide human moderators to nominate effective features for distinguishing approved posts from removed posts. This would be followed by the use of machine learning techniques that weigh these features and produce models that have good accuracy as well as recall and that use human-understandable features. My findings indicate that moderators would appreciate the ability to understand outputs based on such features. As Cheng and Bernstein suggest (Cheng and Bernstein, 2015), the performance of these models could be further improved by identifying spaces where misclassifications occur. Moderators could be asked to nominate additional features that may be informative in improving performance in those spaces. Researchers and practitioners could build, deploy and test such systems on social media platforms, and compare their performance with existing regulation mechanisms.

5.4.4 Limitations and Future Work

This study has some limitations. My results are from interviews with a small sample of Reddit moderators. I note a self-selection bias: I only spoke with Reddit moderators who were willing to talk to me. My sample is diversified in that my participants host a variety of subreddits, come from various geographic areas, and have different occupations. Participants not only moderate the five subreddits that I sampled from, but also a number of other subreddits (Table 5.2). Nevertheless, my sample was all male. I suspect this is because Reddit moderators are disproportionately male; still, to my knowledge, no comprehensive data on the demographics of Reddit moderators exist. Prior research has shown that gender shapes individuals' conceptions of offense in online posts (Binns et al., 2017). Besides, many scholars have studied the issues of gender equity in online forums and their effects on democratic discourse (Herring, 2000; Martin, 2015; Pierson, 2015; Wilburn, 1994). Therefore, future work on the demographics of Reddit moderators and especially how gender affects moderation practices would be a useful contribution.

As I progressed through my interviews, I began to hear the same themes again and again. My final few interviews generated limited new insights, suggesting my data reached empirical saturation. This supports the validity of my results. Additionally, I note that social desirability bias might have affected what my participants were willing to share with me.

One rich direction for future work is to evaluate the performance of Automod and characterize its false positives and false negatives. Moderators could be asked to code whether they would allow or remove a sample of postings on their subreddit, and these codes could be compared with the actual outcomes of Automod processing of those postings. This could provide useful insights into the net workload saved because of the use of Automod and the amount of new workload generated because of the false positives of Automod's decisions. Additionally, Automod's performance could be compared with the results of machine learning models trained on previously moderated data from the subreddit. Analyzing posts on which moderators disagree or find it difficult to take a decision could also provide valuable insights about moderation.

Finally, this study presents the point of view of moderators. In future work, it would be beneficial to study the perspectives of participants whose comments and posts may or may not be deleted by Automod. More generally, analyzing the effects of adopting automated moderation tools on the design of posting guidelines and the demands on end-users on different platforms would provide valuable insights. It would also be useful to investigate when platforms' interests align with and differ from those of volunteer moderators.

5.5 Conclusion

This study presents a qualitative inquiry of the content regulation ecosystem on Reddit, one of the most popular social media platforms. My findings show that content regulation on Reddit is a socially distributed endeavor in which individual moderators coordinate with one another as well as with automated systems. I discuss the benefits of automated tools and identify areas for improvement in the development and sharing of these tools.

I see this study speaking to multiple stakeholders, such as creators of platforms, designers and researchers interested in automated or machine learning-based content regulation, scholars of platform governance, and content moderators. In conclusion, I highlight the following takeaways from this work:

5.5.1 For creators of new and existing platforms

As user traffic increases, moderation systems have to process content at massive levels of scale. This makes it necessary for platforms to use automated tools. My findings suggest, however, that the use of automated tools has direct and secondary effects on multiple stakeholders and their activities — from how moderators coordinate among one another and create community guidelines to how users are required to craft their posts. I therefore recommend that platforms carefully reflect on the anticipated ripple effects over different stakeholders when determining which automated tools they deploy in content regulation systems.

Usually, how content regulation occurs on social media platforms remains a trade secret and is not revealed publicly. In this research, I provide details of how Reddit moderators distribute the work of content regulation between human workers and automated tools. My comprehensive description of Reddit regulation provides an important reference point for how human-machine mixed initiative regulation systems can be designed and deployed on other platforms.

5.5.2 For designers and researchers interested in automated content regulation

I highlight the concerns that designers of machine learning based content regulation should take into account when creating new tools. I found that although Automod relies on syntactic rules instead of advanced machine learning techniques, moderators value Automod because it provides them a great level of control and understanding of the actions taken by Automod. My findings reveal that moderators who do not understand how automated tools work may not be able to contribute as much after these tools are adopted. This can, in turn, affect the dynamics of relationships among the moderator team. This highlights the significance of creating tools whose configurations are easily understood by the moderators, and designing tutorials that assist this understanding.

Furthermore, it is important to explore how the use of automated tools shapes the explainability of moderation decisions and the perceptions of affected users (Clément and Guitton, 2015; Jhaver, Karpfen, and Antin, 2018). I also hope to see studies that investigate how the use of automated mechanisms differs between Reddit and platforms that rely on commercial content moderation firms.

5.5.3 For scholars of platform governance

In recent years, researchers have begun asking questions about the democratic accountability of platform companies and their role in the realization of important public values like freedom of expression, transparency, diversity, and socio-economic equality (Gillespie, 2018a; Gorwa, 2019; Helberger, Pierson, and Poell, 2018; West, 2018; Suzor et al., 2019). My findings contribute to this conversation by showing that an increased reliance on automated moderation tools can contribute to situations where content moderation may seem unfair. Since automated tools can't always consider the context of a post, they may consistently censor individuals with certain viewpoints, or they may influence the discursive norms in unforeseen ways and increase online polarization (Binns et al., 2017). On the other hand, these tools may catch and remove posts that are problematic only at the surface level but allow proliferation of bigoted viewpoints that are subtle and avoid automatic detection at deeper levels of meaning. Exactly how the adoption of various types of automated moderation tools affects different user groups is a subject that scholars of platform governance must examine so that they can articulate strategies that may address the problems of tool bias.

Automated moderation tools not only exacerbate biases but they also operate simply by reacting to problems, not by dealing with their root causes. Such an approach simply hides problematic behaviors such as sexism and racism instead of interfacing with offenders in meaningful ways (Hughey and Daniels, 2013). This may merely push the offensive users to other platforms where their bigoted views are more welcomed. In this way, current automated tools miss out on the opportunities to examine the social and psychological factors that lead to hateful discourses. Instead, I call for researchers to find ways, which may go well beyond simply a deployment of automated tools, to change offensive or uninformed users' perspectives on socially relevant issues (Jhaver, Vora, and Bruckman, 2017; Seering et al., 2019a) and help shift norms in positive ways.

5.5.4 For content moderators

My findings point out the new challenges that moderators can expect to grapple with as they adopt new automated tools in their work. Content moderation is *hard*, and even without the use of automated mechanisms, when moderators update their policies, they sometimes have to revisit previous positions (Gillespie, 2018a). However, delegating content regulation to automated tools increases the distance between how community guidelines are conceptualized by moderators and how they are enforced in practice by automated tools. This distance raises the possibility that moderation systems will make mistakes. Consequently, moderators can expect to take on the additional tasks of correcting the false positives of these tools.

The use of automated tools not only adds to the moderators' work by requiring them to reverse mistakes made by automated moderation, it may also affect the relationship between the users and the moderators (Squirrell, 2019) because of the higher occurrences of mistakes. Using these tools also results in increased user complaints that moderators have to respond to, which further adds to the moderators' work. Besides, moderators may not have control over what is disclosed about the operation of these tools to the ends-users. Although the extent of such disclosures may instead be determined by designers or site administrators, it may still shape end-users' perceptions of moderators' work because end-users may consider moderators responsible for the choices of transparency in moderation decisions. Moderators may also need to develop technical expertise to use advanced machine-learning based tools efficiently. In sum, the use of automated tools changes the work required of moderators and their relationships with end-users in important ways. As community managers inevitably move towards adopting more automated tools for content regulation, efforts to prepare moderators for such changes will be vital.

CHAPTER 6
UNDERSTANDING USER REACTIONS TO CONTENT REMOVALS: A SURVEY OF MODERATED REDDIT USERS

"I feel sad that my effort in making that post was for nothing, and that no one will see it and no one will reply with any help or advice." - P254

6.1 Introduction

How do users feel when their content is removed from online communities? Does it deter them from posting again? Does it change their attitude about the community? Individuals have a range of motivations for posting, and this shapes their reactions to content removal. In some cases (like P254 above), a user might really need advice. In others, as we will see in this chapter, a user might annoy the moderators on purpose, intending to provoke a removal. How does the level of effort made in creating content affect the way users perceive its removal, and does receiving an explanation of why content was removed matter? In this chapter, I address these questions through a survey of 907 Reddit users whose posts were removed[1].

This chapter is concerned with understanding content moderation from the perspectives of end-users in cases where the user has likely broken a rule or a community norm. I focus specifically on content moderation processes that determine which user-generated content to allow on the site and which to remove, as well as how to handle removals. The goal of this research is to offer theoretical and practical insights for community managers to moderate their communities in ways that are considered fair by the end-users and that encourage users to continually submit constructive contributions.

[1] Findings from this study were published in 2019 in the ACM CSCW conference(Jhaver et al., 2019b). Darren Scott Appling assisted me and my PhD advisors, Amy Bruckman and Eric Gilbert, guided me on this work.

In recent years, the fields of Human-Computer Interaction (HCI) and Computer-Supported Cooperative Work (CSCW) have actively engaged in research on various aspects of content moderation, highlighting the differences in moderation policies across various social media platforms (Pater et al., 2016) and exploring the design challenges of creating automated tools for enacting efficient content regulation (Chapter 5). However, much of the research in this area is theoretical and existing empirical work usually takes a data-centered (Chancellor, Lin, and De Choudhury, 2016; Chandrasekharan et al., 2017a; Lampe et al., 2014) or moderator-centered (Matias, 2016c; McGillicuddy, Bernard, and Cranefield, 2016; Seering et al., 2019b) perspective. Limited prior research has investigated the perspectives of moderated end-users on Twitter and Facebook (West, 2018), but to the best of my knowledge, no prior work has explored how users react to content removals on sites like Reddit that rely on volunteer, community-driven moderation. I seek to address this gap in research with my analysis of moderated Reddit users using large-scale survey data.

End-users are the central actors in online social systems. Sites like Reddit don't usually create their own content. Instead, they rely on a constant stream of user-generated content (Gillespie, 2018a). Therefore, end-users are not just consumers who bring in the ad revenue to sustain these platforms but they are also the content creators. As a consequence, it is crucial for these platforms to have users who are invested in the online community and who feel valued for their content contributions.

Although many users on these platforms create information goods that are appreciated by the community, there are others whose posts are promptly removed by the community managers before they can be seen by the community[2]. We do not know what happens to individual users after they invest time in creating content only to have it discarded. It is also unclear how the different elements of submission process (e.g., the existence of community

[2] I do not take the view that all content removal decisions on social media sites are legitimate. For example, automated moderation tools may miss the contextual details of the post and remove it based on the presence of specific keywords in the post (Chapter 5) or human moderators may be motivated by personal biases (Chapter 4). I did not (and as an individual external to the community, and therefore, unaware of its social norms, could not) independently verify whether the removals of my participants' posts were legitimate. Instead, my analysis focuses on whether or not end-users perceive these removals as fair.

guidelines) and the subsequent removal process (e.g., whether or not the user is provided a removal reason) affect users.

Understanding the concerns and experiences of such users may open up opportunities for identifying and nurturing users who have the potential to become valuable contributors in the community. From a wellness perspective, the effects of content moderation on the levels of stress experienced by millions of end-users is also important to consider (Mark, Wang, and Niiya, 2014). Therefore, it is imperative that we study the users who experience content removals.

In particular, this chapter focuses on moderated users' perceptions of fairness in content moderation and how content removals shape their attitude about posting in the future. I consider these two factors as important to understanding the users' current orientation and future outlook towards online communities. My analysis is guided by the following research questions:

- RQ1: How do users perceive content removals?
- RQ2: In what ways do the contextual factors of post submission and content moderation, such as community guidelines and removal explanations, shape users' perceptions of fairness of content removals?
- RQ3: How do these contextual factors affect users' attitude about posting in the future?

To answer these questions, I conducted a survey of users (N=907) who have experienced content removals. I chose to conduct this study on Reddit. As highlighted earlier in this thesis, Reddit communities show a variety of approaches to enacting content curation. For example, some communities display a set of community guidelines or subreddit rules that users should follow while others don't provide any such guidelines (Fiesler et al., 2018). To take another example, moderators on some communities provide explanations for post removals while others choose to silently remove posts. Therefore, the Reddit platform provides a rich site to study how the differences in the contexts of post submissions and subsequent moderation actions affect the attitudes of users.

I triangulate quantitative and qualitative data from this survey to present a rich overview of moderated users, their concerns, and their interactions with various elements of the moderation process. As might have been predicted, a majority of my participants expressed negative attitudes about their content removals. However, analyzing their dominant affective responses and interactions with moderation processes reveal insightful nuances. For example, my qualitative data indicate that the absence of notifications about removals results in users creating folk theories of how content moderation works, and this reinforces their negative attitudes. My quantitative analyses show that having community rules and receiving removal explanations are associated with users perceiving the content removal as fair and having better attitudes about future posting.

This work sheds light on the needs of users whose content has been removed. I add support to prior research that calls for taking an educational approach rather than a punitive approach to content moderation (West, 2018). I offer recommendations for designers to build tools and community managers to adopt strategies that can help improve users' perceptions of fairness in content moderation and encourage them to become productive members of the community. Since moderation resources are often scarce, my empirical findings can help moderators make informed choices about where to invest their effort related to providing community guidelines, offering removal explanations, and using automated tools.

6.2 Study Context: Content Removals on Reddit

For the purpose of this study, I focus on moderation of only the submissions (and not the comments) on Reddit. When moderators remove a submission on Reddit, the site doesn't automatically notify the author of the submission about the removal. While on some communities, moderators explicitly inform the poster that their submission has been removed, most communities choose not to inform the poster. When signed in, posters can still access all their submissions they have posted on their user profile page, regardless of whether they have been removed on Reddit. Therefore, posters may only come to know about the removal if they check the front page of the subreddit and do not notice their post.

When a submission is removed, moderators on some communities also choose to provide posters an explanation of why the removal occurred. This can be done in a number of different ways. Moderators can (1) comment on the removed post with a message that describes the reason for removal, (2) flair[3] the removed post, or (3) send a private message to the submitter. Moderators can either choose to post the removal explanation themselves, or they can configure automated tools (e.g., AutoModerator (see Chapter 5)) to provide such explanations when the submission violates a rule.

In this study, I evaluate how these different design mechanisms mediate user responses to content removals on Reddit.

6.3 Methods

I designed a survey to directly ask users who receive different types of moderation responses questions about their experiences (see Appendix B). I used a modified form of an experience sampling approach (Scollon, Prieto, and Diener, 2009) to collect my participants' responses right after their posts got removed on Reddit. The survey contained 24 questions (mostly multiple choice, with a few free-response). My goal in this analysis was to gain a deep understanding of how variations in moderation affect users. I studied how the users perceive the feedback mechanisms (e.g., subreddit rules, comment describing the reason for removal) that are implemented by the moderators. Most importantly, I investigated how the users' experiences with content removal shape their attitudes about fairness in content moderation and future interactions on the community.

[3] Flairs are short tags that can be attached to users' submissions. Only the moderators on each subreddit have access to assign removal explanation flairs to the posts on that subreddit.

6.3.1 Survey Instrument

The survey questions were based on the tension points around content moderation that have surfaced in prior work (Fiesler et al., 2018; Matias, 2016c; Matias, 2016b) and workshop discussions (Blackwell et al., 2018b; Bruckman et al., 2018; Matias, 2018). To increase the validity of the survey, I conducted an in-person cognitive pretest of the survey with four students at Georgia Tech. These students were not involved with the project and they provided feedback on wording of the questions and survey flow, which I incorporated into the final survey design. I also piloted the survey with a small subset of the sample (28 participants). During this field test, I included this question in the survey: "Q - This survey is currently in pilot stage. Do you have any suggestions for how we can improve this survey?" These survey pretests resulted in several rounds of iteration before my questionnaire reached the desired quality.

The questions in this survey measured the attitudes and perceptions of users concerning posts they made that had recently been removed on Reddit. I asked users how they perceived the fairness of the post removal. My questions captured users' awareness and impression of different features of Reddit moderation system such as subreddit rules and removal explanations. I also included open-ended feedback questions in this survey to understand the relative frequency of key satisfactions and frustrations with moderation systems. These questions asked users: "Please explain how you felt about the removal" and "Is there anything else you'd like to tell us about your view of this removal?"

My questionnaire used skip logic, i.e., I asked a different set of questions to different respondents based on their answers to previous questions. For example, I first asked users whether they noticed any rules on the subreddit they posted to. Only if the participant answered 'yes' to this question did I ask them follow-up questions about whether they read the rules and whether the rules were clear. This was done so as to remove questions that may be irrelevant for some respondents and reduce the time they need to complete the survey. I used Google Forms to implement this survey.

Following the guidelines described in Müller, Sedley, and Ferrall-Nunge (2014), I took several steps to avoid the common questionnaire biases in designing my survey instrument. For example, to minimize satisficing bias, I avoided questions that required an excessive amount of cognitive exertion. To minimize social desirability bias, I allowed participants to respond anonymously (Holbrook and Krosnick, 2009). However, I asked participants to submit their Reddit username so that I could later merge survey responses for each participant with their behavioral data obtained using the Reddit API via the username identifier. To minimize question order biases, I ordered questions in a funnel approach, i.e., from broad to more specific. Earlier questions were easier to answer and were more directly related to the topic of the survey whereas sensitive questions (e.g., about education levels and age) were placed towards the end of the survey so as to build rapport and avoid early drop-off (Dillman, 1978). I grouped related questions together to reduce context switching, and I presented distinct sections on separate pages of the survey for easier cognitive processing. Furthermore, I avoided including broad questions, leading questions or double-barreled questions in this survey (Müller, Sedley, and Ferrall-Nunge, 2014).

I took several steps to discourage disruption from those who might seek to manipulate the data. First, I recruited my participants through private messages instead of publicizing the link to my survey webpage on public forums. Second, I made most questions “required” so that the survey could not be completed until a response was submitted for each question. Third, following the method implemented by West (2018), I adopted a page-by-page design so that users had to click through multiple pages of content in order to complete the survey. I also asked participants to describe in their own words how they perceived the content removal. Although dedicated actors may still manipulate the data, these measures were intended to act as disincentives to providing falsified data in large quantities.

This research was approved by the Georgia Tech Institutional Review Board (IRB). Subjects were not compensated for their participation.

6.3.2 Data Collection

My sampling frame consisted of all Reddit users whose recent post(s) was removed. To minimize selection bias, I used a random sampling approach to select participants: I randomly drew the sample from users in my sampling frame and invited every user in the sample in the same way (Müller, Sedley, and Ferrall-Nunge, 2014). My strategy for determining when to contact users in this sample was motivated by two goals:

1. Enough time should have passed after a submission is posted for human moderators to review and in some cases remove that submission or allow moderators to reverse the removal decisions incorrectly made by automated tools.

2. Submission should be posted recently so that the submitter would easily recall the circumstances around the posting and provide appropriate responses to the questions asked in the survey.

Achieving both of these goals required attaining a balance between sending the survey request soon enough so that the users would have a sufficient recall of their submission but not so soon that moderators haven't had enough time to review that submission. For the purpose of this survey, I configured my data collection such that at least three hours elapsed after the time of submission so that the moderators had enough time to review it. At the same time, I only collected data against submissions that were posted less than 12 hours ago. This selection of time bounds for contacting participants was not based on any empirical tests but it was based on my experiences as a Reddit moderator on a variety of subreddits over the past four years. Subreddits usually have multiple moderators, often located in different time zones, and they review submissions much more promptly than they review comments in order to avoid community engagement with submissions that are subsequently removed.

I began by collecting a random sample of 10,000 (allowed as well as removed) sub-

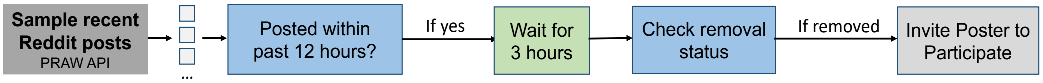


Figure 6.1: Flowchart depicting the data collection process. I created a Python script to automate this process.

missions recently made on Reddit using PRAW Reddit API[4]. This API does not provide a facility to directly retrieve 10,000 random submissions across all subreddits. Therefore, to collect my data, I started with randomly sampling a subreddit and then retrieved the most recently posted submission on this subreddit. I stored this submission if it was posted in the past 12 hours. Next, I repeated this process until I got 10,000 random submissions posted in the past 12 hours (Figure 7.1).

After I retrieved the 10,000 submissions in the previous stage, I waited for three hours so that the moderators had sufficient time to review and moderate those submissions. At the end of this waiting period, I tested the removal status of the collected submissions, again using PRAW Reddit API. Next, I retrieved the authors of submissions that were removed and sent them a customized invitation message to participate in the survey, with each message containing the link to corresponding user's removed submission. I created a Python script to automate this entire process, which included using Reddit API to send out a customized survey invitation to moderated users through a Reddit private message. Using this script allowed me to ensure that the messages were promptly sent to all participants.

Because I randomly sampled subreddits in this process, regardless of their popularity or subscriber size, my sampling strategy allowed me to have a broad coverage of subreddits in my sample. It was important for me to have this diversity of communities and to not have my data representing only a few, large subreddits because I wanted to measure users' responses to a wide range of moderation practices in my survey (Appendix B). Yet, since I selected only those posts that were submitted in the past 12 hours, my sampling favored subreddits that were at least somewhat active.

[4] https://praw.readthedocs.io/en/latest/

I repeated this process for seven days at which point I got the target number of responses to my survey. 8.2% of all users I contacted submitted my survey. I considered this to be a suprisingly high response rate given that I sent my survey requests through online private messages. I attribute this high response rate to the customized private messages I sent using my Python script. Many of my invitees replied back to my invitation message with clarifying questions about the survey. I answered such questions as often as I could, and I believe, this also helped boost the survey response rate. I ensured that I didn't send invitation message to any user more than once. Furthermore, I initiated my Python script for data collection at different times of the day everyday so that users posting at different times and in different time zones could be selected.

6.3.3 Data Preparation

After the data collection was completed, I proceeded with data preparation and cleaning. First, I removed all duplicate responses by looking for multiple entries that contained the same username. As mentioned before, my data collection was configured such that each user was sent only one invitation to respond to the survey. Therefore, my records contained a unique subreddit for each user where the post removal occurred. My survey asked respondents to type in the name of the subreddit that their removed submission was posted to. This was an attention check question. I manually matched the answer posted to this question against my records to verify that the participant was responding about the removed submission I had on my records, and removed all survey responses where a mismatch occurred.

I read all responses to the open-ended questions in this survey and manually removed the obvious garbage answers such as "abcd." I also examined other answers from the same respondent to determine whether all answers from that respondent warrant removal (Müller, Sedley, and Ferrall-Nunge, 2014). It is, of course, possible that participants were not truthful about their experiences of content removals. I acknowledge, however, that response bias is inherent in any self-report dataset obtained, and my results should be interpreted with this in mind.

6.3.4 Participants

In total, 1,054 users clicked through the consent form and submitted the survey. I filtered out inappropriate survey responses using the data cleaning process described in the last subsection. My final sample consisted of 907 responses. Participants came from 81 different countries, although there was a heavy skew toward North America. The four countries with the most respondents were the U.S. (61%), Canada (7%), U.K. (5%), and Australia (3%). Seven respondents elected not to provide their place of origin. A majority of the participants were male (81%) and under 25 years old (55%). Comprehensive demographic information is reported in Table 6.1.

I used the Reddit API to get additional information about my participants related to their activity level on Reddit. Participants had a median of 3,412 karma points (μ = 45K, σ = 421K). Their account age had a median of 436.6 days (μ = 804.93, σ = 496.6). Participants had posted a median of 35 submissions on Reddit (μ = 258.06, σ = 1195.7).

6.3.5 Variables

The survey asked respondents about their agreement with the statements (1) "I think that the removal was fair" on a five-point Likert scale (1=Strongly Disagree, 5=Strongly Agree), and (2) "How likely are you to post again on this subreddit after this experience?" also on a five-point Likert scale (1=Very Unlikely, 5=Very Likely). I used the answers to these questions as dependent variables in each of my regression analyses (Table 6.2). I refer to these variables as *Fairness* and *PostAgain*, respectively through the rest of this paper. I note here that the variable *PostAgain* measures my participants' *perception* of their likelihood of future posting, and not whether they actually posted again on the subreddit. While analyzing the actual future behavior of moderated users is an important research direction,

Table 6.1: Participant demographics (N = 907). Note that the percentage of each factor does not always sum to 100% because of rounding. A majority of participants were from North America. The most frequent age group was 18-24. 81% of participants identified as male.

Factor	Category	% (N)
Country	United States	61% (554)
	Canada	7% (68)
	United Kingdom	5% (46)
	Australia	3% (31)
	India	2% (14)
	Other	21% (187)
	Prefer not to Answer	0.8% (7)
Age	18-24	55% (502)
	25-34	23% (210)
	35-44	11% (100)
	45-54	5% (45)
	55-64	1% (10)
	>65	0.3% (3)
	Prefer not to Answer	4% (37)
Education	Less than High School	15% (133)
	High School	18% (160)
	Some College, No Degree	20% (185)
	Bachelor’s Degree	22% (200)
	Master’s Degree	9% (81)
	Associate degree	6% (57)
	Doctorate Degree	2% (18)
	Prefer not to Answer	8% (73)
Gender	Male	81% (738)
	Female	13% (121)
	Another Gender	1% (14)
	Prefer not to Answer	4% (34)

Table 6.2: Control variables (including demographic and prior history variables), independent variables (including post context, community guidelines and removal explanations variables) and dependent variables included in my regression analyses.

Control Variables
Demographics
(1) Age
(2) Education
(3) Gender
Prior History
(1) Reddit karma
(2) Time on Reddit (in days)
(3) Number of submissions posted on Reddit

Independent Variables
Posting Context
(1) Time spent in creating submission
(2) Suspecting removal before posting
(3) Noticing post removal
Community Guidelines
(1) Reading the rules
(2) Understanding the rules
Removal Explanations
(1) Explanation providing new information
(2) Explanation mode
(3) Explanation source

Dependent Variables
(1) *Fairness*
(2) *PostAgain*

and I pursue that direction in Chapter 7 with a larger dataset[5], the current chapter focuses on understanding users' attitudes. Therefore, using the *PostAgain* dependent variable, I seek to explore users' beliefs about their future posting just after they experience content moderation.

For each user, I gathered features that I hypothesized would be related to my dependent variables. These features can broadly be categorized into three different buckets: (1) Posting context, (2) Community guidelines, and (3) Removal explanations. Table 6.2 shows the list of variables I included in my models.

I included the following features that related to the context of the posting as independent variables in my analyses: (a) The amount of time spent in creating submission, (b) whether the participant suspected the post would be removed before submission, and (c) whether the participant noticed the post removal before starting the survey. Through these variables, I wanted to test whether the amount of time users spend in composing their posts has an effect

[5] Citation omitted for anonymous review.

on their attitudes about content removals. I was curious to see whether the participants who did not expect a post removal before submission have different reactions to content removals than users who suspected a removal. Finally, I wanted to analyze whether the users who were not notified about the removal of their post and only came to know about it through my survey request have different reactions to content moderation than users who were notified about the removal. I test for these associations in my regression analyses.

In the context of community guidelines, I used responses to two questions on a Likert scale as my independent variables: (a) "I read the rules of the subreddit before posting," and (b) "The rules on this subreddit are clear." Limited prior research shows that highlighting the rules has an effect on improved user behavior (Matias, 2016b). I wanted to test whether reading the rules has an association with participants' attitudes as well. I also seeked to study whether the clarity of rules improves users' perceptions.

As I discussed in Section 6.2, on Reddit, moderators can provide explanations for post removals in a variety of ways. They can comment on the removed post, flair the post, or send a private message to the poster. Moderators can either compose these messages themselves or they can configure automated moderated tools to automatically send explanations whenever a submission matches a known pattern. Given this context, I wanted to test these factors for their associations with my dependent variables: (a) Whether the explanation provided new information to the participant, (b) Source of explanation ('human', 'bot' or 'unsure') and (c) Mode of explanation ('comment to submission', 'private message' or 'flair'). Testing for these associations allowed me to explore how the ways of providing explanations and the novelty of information in explanations affect users' attitudes.

I selected these independent variables because they were open to direct interpretation and I wanted to test their relationships with attitudes about fairness in moderation and future postings. My selection of these variables as factors of interest was based on intuitions developed through serving as a moderator and content contributor on many Reddit communities over the past four years. I note that my analytic approach is designed to be primarily exploratory, as these constructs have not yet been examined in literature.

In each of my models, in keeping with prior literature on perceptions of social media users (Saha, Weber, and De Choudhury, 2018; Saha et al., 2019), I used the participants' demographic characteristics (age, education level and gender) and prior history on Reddit (as measured by their karma score, number of days since they created their account, and number of submissions on Reddit) as control variables (Table 6.2). I treated gender as a nominal variable and age and education as ordinal variables in my regression analyses. Further, I treated all the 'Prefer not to Answer' entries as missing values in my models. I note that although the control and independent variables discussed above capture many important factors that may mediate user reactions to content removals, there are other factors related to user demographics (e.g., race) and subreddits (e.g., topic) that I do not control for in my analyses. Therefore, I see my models as reasonable simplifications of the Reddit sociotechnical system.

6.3.6 Data Analysis

I used linear regression models for their easy of interpretability after checking for the underlying assumptions. I created separate regression models for evaluating the effects of posting context variables, community guidelines variables and removal explanations variables (Table 6.2) on my dependent variables. I evaluate separate models because given the skip-logic nature of my questionnaire, only certain subsets of participants had responses to some lines of questions (Appendix B). In addition to these analyses, I also conducted separate tests to understand the associations of noticing community guidelines and receiving removal explanations with my dependent variables. When building each regression model, I performed listwise deletions of the cases where any of the input variable value was missing.

For the open-ended questions, I iteratively developed a set of codes based on an inductive analysis approach (Straus and Corbin, 1998). I coded for these questions together

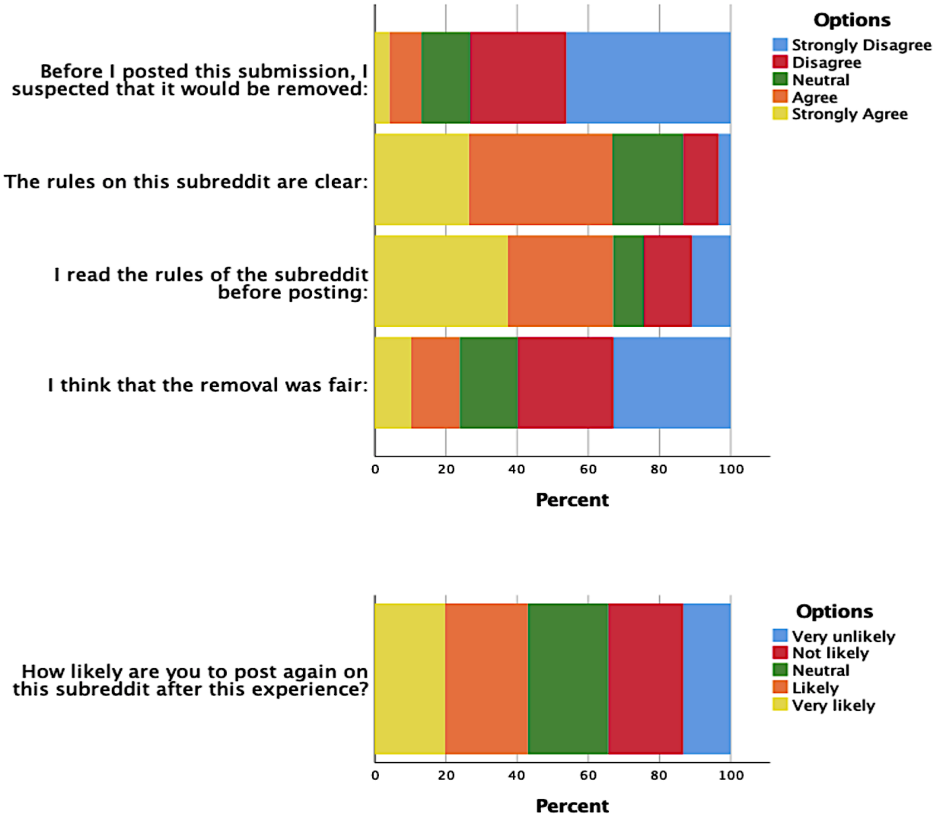


Figure 6.2: Frequency of participants' responses to various survey questions, measured in percentage.

because each question had responses that pertained to themes about perceptions of content moderation. This process resulted in a codebook with ten codes (Table 6.7). My coauthor, Scott Appling and I coded all open-ended responses side-by-side in order to iterate on the codes and to double-check the results. All disagreements between the two coders were resolved through discussions.

6.4 Quantitative Findings

I first calculated the descriptive statistics for the dependent variables *Fairness* and *PostAgain*. Overall, I found that 10.3% of all participants strongly agree and 13.6% agree that the removal was fair whereas 33% of participants strongly disagree and 26.9% disagree that

Table 6.3: Regression analyses for (1) whether users perceive the removal as fair (*Fairness*) and (2) whether users are likely to post again on the corresponding subreddit (*PostAgain*). This model includes posting context variables as independent variables in addition to the control variables as inputs.

		Fairness				***PostAgain***			
		B	**SE**	β	***p***	**B**	**SE**	β	***p***
	Intercept	1.857	0.219		.000	3.606	.240		.000
Control Variables	**Age**	-0.039	.052	-.029	.451	-.016	.057	-.012	.780
	Education	-0.041	.030	-.052	.169	.005	.033	.007	.867
	Gender	0.165	.130	.043	.204	-.252	.142	-.066	.076
	Reddit Karma	1.3E-8	.000	.004	.921	-6.2E-9	.000	-.002	.965
	Reddit Age	0	.000	.073	.041	4.5E-5	.000	.031	.431
	Submission Count	2.5E-5	.000	.021	.630	8.3E-5	.000	.071	.140
Posting Context Variables	**Post Creation Time**	-0.18	.050	-.118	.000	-.203	.055	-.136	.000
	Noticed removal	0.362	.045	.265	.000	.051	.049	.038	.297
	Expected removal	0.396	.039	.334	.000	.096	.043	.083	.025

the removal was fair. 16.3% of participants felt "neutral" about the fairness of moderation (Figure 6.2). I also found that 19.7% of all participants considered it very likely and 23.3% considered it likely that they will post again on the subreddit where their submission was removed while 13.3% of participants felt it very unlikely and 21.1% considered it unlikely that they will post again. 22.6% of participants felt "neutral" about this factor (Figure 6.2).

In the rest of this section, I explore how these attitudes are linked to different aspects of user experience on Reddit such as posting context, submission guidelines and removal explanations.

6.4.1 Posting Context

To identify how posting context is associated with my dependent variables, I conducted a linear regression for each of the two dependent variables. I used posting context variables as independent variables in these analyses. Table 6.3 shows the results of these regressions. My baseline models, that used only the control variables (n=738), explain only 0.6% (adj. $R^2 = .006$) and 0.8% (adj. $R^2 = .008$) of variance in *Fairness* and *PostAgain* respectively (Table 6.4). Adding the posting context variables (n=738) increases the adjusted R^2 values to .219 and .033 respectively.

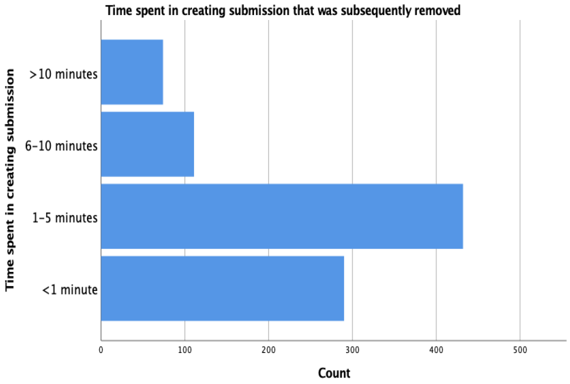


Figure 6.3: Time spent in creating submissions that were subsequently removed, as reported by participants.

I asked participants how much time they spent in creating their submission. My survey data show that 31% of participants took less than a minute and only 8% of participants spent more than 10 minutes to create their submissions (Figure 6.3). As Table 6.3 shows, time spent in creating submissions is significantly related to both *Fairness* and *PostAgain* even after controlling for other factors. Results indicate that as time spent increases, participants are less likely to consider the removal as fair (β = -.118) and less likely to consider posting in the future (β = -.136). One possible explanation for this is that users who spend more time crafting their post are more invested in their post, and therefore, they feel more aggrieved at the removal and less motivated to post again.

Next, I explored how often the users even notice it when the content they post on Reddit is removed. To my surprise, 41.8% of my respondents (n=379) reported they did not notice that their post was removed until they received my invitation message to participate in the survey. I received dozens of replies to my invitation messages where users expressed surprise about their post removals. Many users I contacted also complained about not receiving any notification about the removals. My regression analyses (Table 6.3) show that when participants notice their removal, they are more likely to consider the removal as fair (β = .265). This suggests that when users become aware of their post removals, perhaps through communication by the moderation team, they are more likely to consider the moderation of their post as fair than when they are not made aware. Noticing removals,

Table 6.4: Summary of regression analyses for (1) whether users will perceive removal as fair (*Fairness*) and (2) whether users consider it likely that they will post again on the subreddit (*PostAgain*). Asterisk levels denote p<0.05, p<0.01, and p<0.001.

	Fairness		***PostAgain***	
Model factors	**Adj. R^2**	**F**	**Adj. R^2**	**F**
Baseline	.006	1.755 (6, 732)	.008	2.007 (6, 732)
Baseline + Posting context	.219	24.06 (9, 729) ***	.033	3.842 (9, 729) ***
Baseline + Community guidelines	.124	10.307 (8, 518) ***	.055	4.852 (8, 518) ***
Baseline + Removal explanations	.044	2.459 (9, 296) **	.000	.989 (9, 296)
Baseline + Posting context + Community guidelines	.298	21.305 (11, 515) ***	.067	4.451 (11, 515) ***
Baseline + Posting context + Community guidelines + Rem. explanations	**.324**	8.279 (14, 212) ***	**.123**	3.262 (12, 212) ***

however, did not have any statistically significant effect on whether participants consider it likely they will post in the future.

I also asked in my survey whether users suspected that their submission would be removed before they posted it on Reddit. My results showed that 73.2% of participants “disagree” or “strongly disagree” that they suspected a post removal whereas only 13.1% of participants “agree” or “strongly agree” that they expected a removal (Figure 6.2). My regression analyses (Table 6.3) suggest that suspicion of post removals before submission is positively associated with *Fairness* ($\beta = .334$) as well as *PostAgain* ($\beta = .083$). This indicates that users who expect a removal prior to posting their submissions are more likely to consider the removal as fair and less likely to be deterred from future posting by the moderation process.

6.4.2 Community Guidelines

On each Reddit community, moderators can provide community members with explicitly stated guidelines called subreddit rules. These rules appear in the sidebar of each subreddit and they describe the injunctive norms[6] of the community. They are one of the key design

[6] Injunctive norms are norms that set behavioral expectations by prescribing acceptable community practices (Cialdini, Kallgren, and Reno, 1991)

elements on Reddit for allowing community managers to encourage voluntary compliance with behavior norms. Kiesler et al. suggest that when social norms are clearly stated through explicit rules rather than being left for users to reasonably infer for themselves, users are more likely to comply with those norms over a variety of situations (Kiesler, Kraut, and Resnick, 2012). However, not all subreddits choose to create and display community guidelines. In a recent analysis of a sample of Reddit communities, Fiesler et al. found that only half of the communities had explicit rules (Fiesler et al., 2018).

I analyzed how users' attention to the presence of these rules in a community affect their attitudes. First, my results show that 71% of participants (n=644) claimed that the subreddit they posted to contained rules in its sidebar, 6.1% (n=55) said there were no rules, and 22.9% (n=208) were unsure. I built a regression model for *Fairness* (n = 738) using the independent variable *containsRules* (this measures whether the participants noticed the rules) and adding the six control variables (Table 6.2). This model showed that noticing the rules was positively associated with perception of the removal as fair ($\beta = .068$, $p < .05$). However, a regression model for *PostAgain* using *containsRules* as an independent variable and including the control variables (n = 738) did not find a significant association for *containsRules* ($\beta = .033$, $p = .432$).

A significant association between *containsRules* and *Fairness* highlights that making rules more prominent may improve users' attitude towards online communities. Thus, creating injunctive norms for the community and nudging users to attend to them is a valuable and underused (Fiesler et al., 2018) moderation strategy that more community moderators should consider adopting.

I asked the participants who noticed the subreddit rules (n=644) two additional questions: (1) whether they read the rules just before posting and (2) whether they perceived the rules of the subreddit to be clear. Results showed that of all the participants who noticed the rules, 66.9% of participants (n=431) “agree” or “strongly agree” and 24.4% (n=157) “disagree” or “strongly disagree” that they read the rules. Moreover, 66.8% of participants

Table 6.5: Regression analyses for (1) whether users perceive the removal as fair (*Fairness*) and (2) whether users are likely to post again on the corresponding subreddit (*PostAgain*). This model includes community guidelines variables as independent variables in addition to the control variables as inputs.

		Fairness				***PostAgain***			
		B	**SE**	β	***p***	**B**	**SE**	β	***p***
	Intercept	2.81	0.32		.000	2.744	0.317		.000
Control Variables	**Age**	-.056	.066	-.040	.403	.029	.066	.021	.664
	Education	-.072	.038	-.087	.061	-.036	.038	-.046	.342
	Gender	.162	.164	.041	.324	-.23	.163	-.062	.159
	Reddit Karma	-1.3E-7	.000	-.049	.357	-8.2E-8	.000	-.032	.558
	Reddit Age	.000	.000	.081	.068	-1.9E-5	.000	-.013	.770
	Submission Count	7.0E-5	.000	.064	.231	.000	.000	.131	.020
Community Guidelines Variables	**Read Rules**	-.308	.041	-.314	.000	-.063	.040	-.067	.119
	Rules are Clear	.222	.054	.170	.000	.278	.053	.223	.000

(n=430) "agree" or "strongly agree" and 13.3% (n=86) "disagree" or "strongly disagree" that the rules are clear (Figure 6.2).

In order to examine the relationships between users' interaction with community guidelines and their perceptions of content moderation, I created linear regression models for *Fairness* and *PostAgain*, including the degree to which the user perceived the rules to be clear and read the rules, as independent variables, and accounting for the control variables (n=526). These models were built using only the responses from the participants who agreed that they had noticed the rules in the subreddit sidebar, and were therefore shown additional questions that pertained to the rules. Each of these models explained a significant amount of variance and had adjusted R^2 value of .124 and .055 respectively, a notable improvement over the baseline models (Table 6.4).

My results (Table 6.5) show that reading the rules was significantly associated with *Fairness* even after controlling for demographic and prior Reddit history variables. As shown in Table 6.5, when users read the rules, they are less likely to consider the removal as fair (β = -.314). This is surprising as one would expect that reading the rules would help users understand the expectations of the community and improve the perceived legitimacy of content removals. However, as I will discuss in Section 6.5, users sometimes have difficulties complying with the subreddit rules because they are too difficult to follow or subjective. Moreover, some users complain that their posts get removed despite compliance with the community guidelines. I did not find a significant relationship between reading the rules and *PostAgain* after accounting for control variables.

I also found that when users perceive the rules to be clear, they are more likely to consider the removal as fair (β = .170) and more likely to consider posting in the future (β = .223) (see Table 6.5). This indicates that clarity of rules has positive association with user attitudes.

In sum, these findings suggest that composing clearly written community guidelines can render the content removals more palatable to the users and motivate them to continue posting despite the current removal.

6.4.3 Removal Explanations

I examined how different aspects of removal explanations affect users' attitudes about content removals and future postings. My results show that in 39.7% of cases (n=360), participants claimed that they were provided an explanation for their removal by the moderation team.

I built a regression model for *Fairness* (n = 738) using the independent variable *receivedExplanation* (this measures whether the participants received an explanation for post removal) and adding the six control variables (Table 6.2). This model showed that receiving an explanation was positively associated with perception of the removal as fair (β = .384, $p < .001$). Similarly, a regression model for *PostAgain* using *receivedExplanation* as an independent variable and including the control variables (n = 738) found that when participants receive an explanation, they are more likely to consider posting again in the future (β = .088, $p < .05$). These results suggest that explanations can be a useful mechanism to gain trust with the users.

I asked users who had received explanations (n=360) additional questions about the explanations, and analyzed the relationships between different aspects of explanations and user attitudes. My results show that 64.2% (n=231) of these participants answered ‘yes’ to the question of whether the explanation provided them any new information, and the rest answered ‘no’.

Next, I found that 57.8% of participants (n=208) received a removal explanation through a comment to their removed submission, 36.1% (n=130) received a private message explaining why their post was removed, and 6.1% of participants (n=22) had their submissions flaired with a short text that explained the post removal. My results also showed that 20% of my participants who received explanations (n=72) felt that their removal explanation was provided by a human moderator, 50.6% of participants (n=182) felt that a bot provided explanation, and the rest (n=106) were unsure.

In order to examine relationships between different aspects of removal explanations and users’ perceptions of moderation, I created regression models that included the mode of explanation (‘comment’, ‘flair’ or ‘private message’), source of explanation (‘human’, ‘bot’, or ‘unknown’) and whether the user perceived the explanation as informative, as independent variables (n=305). These models were built using only the responses from the participants who agreed that they were provided an explanation about the post removal, and were therefore shown additional questions related to explanations. As in earlier models, I used *Fairness* and *PostAgain* as dependent variables and demographic and prior Reddit history variables as control variables. Table 6.6 shows the results of these analyses.

Results show that the only explanation factor that explained a significant amount of variance in the perception of removal as fair was considering the explanation as informative (Table 6.6). When users feel that explanations provide them information that is novel, they are more likely to perceive the removal as fair (β = .200). This is possibly indicative of cases where users mistakenly violate a rule or community norm they weren’t aware of when submitting a post and realize their mistake when receiving explanations. However, perceiving the explanation as informative did not have a significant association with *PostAgain*.

Table 6.6: Regression analyses for (1) whether users perceive the removal as fair (*Fairness*) and (2) whether users consider it likely that they will post again on the subreddit (*PostAgain*). This model includes removal explanation variables as independent variables in addition to the control variables as inputs.

		Fairness				***PostAgain***			
		B	**SE**	β	***p***	**B**	**SE**	β	***p***
	Intercept	2.895	.379		.000	3.332	.382		.000
Control Variables	**Age**	-0.076	.096	-.053	.433	.077	.097	.054	.431
	Education	-0.095	.054	-.116	.077	.066	.054	.081	.227
	Gender	0.299	.225	.078	.185	-.143	.227	-.038	.530
	Reddit Karma	-2.3E-9	.000	.000	.998	5.9E-7	.000	.047	.524
	Reddit age	.000	.000	.174	.005	2.2E-5	.000	.014	.823
	Submission Count	4.7E-5	.000	.040	.588	4.5E-5	.000	.038	.609
Removal Explanation Variables	**Explanation Provides New Information**	0.286	.081	.200	.000	.051	.082	.036	.535
	Explanation mode	0.008	.084	.005	.925	-.089	.085	-.063	.295
	Explanation source	-0.045	.111	-.023	.683	-.054	.112	-.028	.627

Neither the explanation mode nor the explanation source had a significant relationship with either *Fairness* or *PostAgain* (Table 6.6). Thus, the mode of explanation does not appear to be important to the users. Moreover, whether a human moderator or an automated tool provides the removal explanation does not seem to matter to the users. Through my experiences as a Reddit moderator, I know that moderators often use pre-configured removal explanations in order to expedite moderation tasks. Thus, the text outputs for both human and automated explanations on many communities look quite general and not specific to the submission at hand. This may be the reason why the users seem to have similar responses to both human and bot explanations.

Comparing the different regression models (Table 6.4), I found that adding either posting context variables or community guidelines variables to the baseline (that includes only the control variables) increases the adjusted R^2 values by substantial amounts for both *Fairness* and *PostAgain*. However, adding removal explanations variables does not contribute as much increase in adjusted R^2 value for *Fairness*, and in fact, the value for *PostAgain* reduces over the baseline model. Combining different sets of variables, I found that the best-fitting models for *Fairness* (adjusted R^2 = .324) and *PostAgain* (adjusted R^2 = .123) were generated by including all the input variables (Table 6.4).

Table 6.7: Kinds of responses mentioned by participants (N=849). Total adds up to more than 100% since many participant responses fit into more than one category.

Theme	Frequency
Lack of clarity about why post was removed	36.9%
Frustration at post removal	28.7%
Perception of moderation as unjust	28.5%
Acceptance of removal as appropriate	18.3%
Frustration at (lack of) communication about post removal	15.7%
Indifference to post removal	12.0%
Difficulties complying with the rules	11.9%
Difficulties with use of automated moderation tools	3.9%
Greater understanding of how to post successfully	3.2%
Satisfaction of removed post receiving many responses	2.5%

6.5 Qualitative Findings

In this section, I report the results of qualitative analysis of open-ended responses to my survey. These results add nuances to the findings presented above in Section 6.4. My participants revealed often negative yet complex attitudes towards content moderation. Table 6.7 summarizes the 849 responses to the open-ended questions about perceptions of content moderation in the survey. I note that the total percentages add up to more than 100% since at times there were overlapping themes that emerged from user statements, and I classified such statements into more than one category.

For the remainder of this section, I use excerpts of quotes from respondents to show representative examples of emergent themes from each coded category in Table 6.7.

6.5.1 Frustration and Lack of Clarity

28.7% of respondents felt frustrated at the post removals. A recurring theme among the responses of these participants is that they felt their efforts at content creation were “not appreciated” on the subreddit they posted to. Many of these participants mentioned being “embarrassed” by the removal while others reported feeling dejected and demotivated from engaging with the community. In line with the relationship I found between time spent in creating submission and users' attitudes through quantitative analysis (Section 6.4.1), participants' open-ended responses reflect that those who had spent considerable time and effort creating a submission felt particularly annoyed about the removals. For example, Participant P893 wrote:

> *"I was confused and angry. It was a good post that I spent half an hour making and there was nothing against reddit's (or the subreddit's) rules! There was no reason to remove the post and honestly, I'm quite furious."*

Participant P254, who stated that she is autistic, and posted on the r/disability subreddit, wrote:

> *"I am autistic and it takes significant effort for me to write up things to communicate effectively and in a way that will be received well. I feel sad that my effort in making that post was for nothing and that no one will see it and no one will reply with any help or advice."*

36.9% of respondents wrote about the lack of clarity in their post removals. A frequent complaint among these respondents was "I have no idea what I did wrong." 64 participants mentioned feeling "confused" about the removal. Many participants argued that they had seen posts similar to their own removed post appear on the subreddit before, so they felt "non-plussed" why their post was targeted for removal. Some respondents pointed out that they were cautious in ensuring that they adhered to all the community guidelines, and yet, their post was removed. Such removals left users uncertain of how they can make successful submissions. For instance, Participant P779 wrote:

> *"I read the rules and my submission was within the guidelines, so I have no idea why it was removed and I'm a little annoyed about it."*

3.9% of respondents complained that their post was mistakenly removed by an automated moderation tool. Some of these participant expressed frustrations about the excessive reliance of moderators on automated tools that often make mistakes.

> *"Mods rely on bots too much. Sometimes there is no human to see why it was removed." - P55*

Content moderation is usually focused on the goal of curating the best possible content. However, given the large proportions of content removals and the frustrations of a majority of moderated users as discussed above, community managers must consider how to assuage the users' frustrations that are linked to post removals.

6.5.2 Perception of Moderation as Unjust

28.5% of participants noted that the moderation was unjust. Some of these users felt that they are unfairly censored despite their adherence to the community guidelines because their posting history indicated an unpopular political affiliation. For example, one participant wrote:

> *"I used to argue with mods but since I participate in some edgy subreddits, lots of mods don't like me and will ignore me, even though I am not rude and my post follows the rules." - P594*

Many participants whose politically charged submissions were removed without notification created their own folk theories about why the removal occurred. Some users felt that the moderators on the subreddit they posted to were politically biased. Others worried that influential online communities often promote a particular worldview and that all the "dissenting voices" are removed. These participants often complained about their inability to exercise their "freedom of speech" and they felt they were being silenced. Some of these participants held that a small number of moderators, who are not elected by the community, had too much power over what gets seen by a large number of users. For example, Participant P23, who did not receive any removal notification, wrote:

> *"It's completely unfair. I didn't break the rules and just got punished for disagreeing with the mods' personal opinions."*

I found that 2.5% of participants justified the validity of their posts by pointing to the positive community response their posts received. They argued that the post removal was unwarranted because other users in the community interacted with the post in a supportive way but the moderators still decided to remove the post. For instance, Participant P815 wrote:

> *"Considering my post had almost 250 upvotes with 99% [up-votes to down-votes] ratio, I'd say everyone in the community enjoyed it and it shouldn't have been removed."*

A few of the participants revealed bigoted generalizations and conspiracy theories about how content moderation happens. For example, one participant wrote:

> *"Criticism of Jews is generally forbidden on Reddit for some reason. It's weird how I can question the teachings of Jesus and the Moon landing and the shape of the Earth, but I must never question the Jew." - P138*

In sum, many users have folk theories of content moderation being shaped by partisan community managers. To what extent these folk theories are accurate and reflect the existing biases of moderators is an interesting empirical question that warrants further research.

6.5.3 Acceptance of Removal as Appropriate

18.3% of respondents indicated an acceptance of their post removal as appropriate. Participants accepted the removal of their posts for a variety of different reasons. Many of these users acknowledged that they had not read the subreddit rules, and they felt that their post removal was valid because they inadvertently violated a subreddit rule. Some of these users expressed regret about not attending more carefully to the community guidelines. For example, Participant P736, who received a private message from the moderators explaining why his post was removed, said:

> *"I felt bad that I had not read the rules and posted inappropriately."*

3.2% of participants mentioned that the content removal helped them become more aware of the social norms of the community and provided them an understanding of how to post successfully in the future. For example, Participant P151 wrote:

> *"I will no longer post images to that subreddit now that I know not to."*

Some individuals explicitly described themselves as "a troll" and they expected that their content would be removed. These users found value in having their blatantly offensive posts be viewed by the community before it is taken down by the moderators. For example, Participant P857 characterized his own post as "disgusting" and he hoped that the post would generate "confused and funny reactions before it was removed." In a similar vein, Participant P375 wrote:

> *"I frequently participate in so called "shitposting" i.e I post content with little to no purpose or meaning behind it. The reasoning behind this is primarily for personal entertainment."*

Another group of users who accepted the removal as appropriate were those who suspected that their post would be removed but they submitted their posts anyway in order to show their group alignments. For example, one participant pointed out that he continues to post inappropriate content on the r/Patriots subreddit, an online community for supporters of the Patriots, an American football team, despite repeated removals because he hates that team. In a similar vein, Participant P90 described his behavior as motivated by a need to prosletyze, and did not feel bothered by the removals of his posts:

> *"Expected, but even if only one person repents, I feel I did what I wanted."*

Thus, participants who accepted the removals as fair include those who realize their mistakes and show an inclination to improve in the future as well as those who have a need to vent on Reddit and did not feel bothered by removals.

6.5.4 Communications about Post Removals

15.7% of respondents complained about the communication, or more frequently, a lack of communication from the moderation team about the post removal. One common sentiment was people reporting frustration about the silent removal of their posts. This reflects the relationship between users not noticing the post removal and perceiving removal as unfair that I found through my statistical analysis (Section 6.4.1). These users often felt cheated upon when they realized that their post was no longer available on the site. For example, Participant P85 wrote:

> *"Whatever I did wrong, the mods should have told me up front. I feel left out of the loop. I probably won't post anything there until I can find out exactly what the problem was."*

Many participants pointed out that they felt frustrated at not receiving any explanations for why their content was removed. This is in line with my findings from Section 6.4.3 that users who do not receive removal explanations are significantly more likely to perceive the removal as unfair. Some users reported being more indignant at the lack of transparency about the moderation process rather than at the removal itself. For example, Participant P838 wrote:

> *"The removal is ok, doesn't bother me, but it's not ok that I didn't get informed."*

A few participants reported dissatisfaction with their interactions with the moderation team about the content removals. Some noted that even after they corresponded with the moderators, they could not understand why their post was removed. For example, Participant P737 wrote:

> *"I sent a polite and thoughtful private message to the mods asking how I could repost or avoid the problem in the future and was given a one sentence*

> *response... Also, the reply did not help me understand the mods' issue with my post."*

6.5.5 Difficulties Complying with the Rules

11.9% of respondents mentioned that they had difficulties complying with the community guidelines. Reddit is a decentralized platform and every Reddit community has its own set of posting guidelines, moderators and social norms. Some respondents who engaged in multiple communities found it tiresome to keep track of and comply with the posting guidelines for each new community they post on. For example, one participant wrote:

> *"I rarely post, in part because each subreddit has a long list of rules and I'm always concerned that I'll miss something, like I did this time. In too much of a hurry? Skimmed the list of rules too quickly? Rules can be difficult to locate when on mobile...There are 2 millions subreddits each with their own list of rules about what you're allowed to say...I just can't be bothered with it anymore." – P902*

Some participants did not understand the reasoning behind why certain rules were put in place on the subreddit. For instance, one participant expressed surprise at finding out that his post was removed because it violated the rule of mentioning the name of a subreddit. This participant was confused about why such a rule was put in place. Other respondents disapproved of certain subreddit rules, and chose to violate those rules deliberately despite suspecting that their post might be removed as a result. For example, one participant pointed out that he posed a question in his submission despite knowing that according to the rules of that subreddit, questions are only allowed to be posted in designated submission threads. He elaborated:

> *"[My submission] technically violates the rules. But having a large thread for questions makes it REALLY easy for things to get lost and never answered,*

> *which happens a lot in subreddits like this. So, I made a thread. Thought it might get removed, and it did...I think a rule that prohibits questions is a terrible idea. If it makes the subreddit spammy, I can understand, but you're stifling community interaction." – P194*

Many respondents complained that it takes a lot of effort to ensure compliance with certain subeddit rules when they post. For example, one participant said that complying with the rule that "a similar link hasn't been posted before" requires putting in a lot of work that involves searching through the prior posts of the subreddit, and it is easier to simply post the submission and hope that it does not get removed. Another participant wrote:

> *"The technical grounds for removal were accurate. However, the rules that qualify a post for removal are overly broad, arbitrarily enforced, and many times onerous to comply with." – P09*

Some participants felt frustrated that the subreddit rules were too subjective and could be interpreted in multiple ways. This mirrors my finding from Section 6.4.2 that users who found the rules to be unclear were more likely to perceive the removal as unfair. A few respondents felt that the moderators deliberately designed unclear rules so that they could defend their removal actions.

> *"Well, my submission was removed because [of the rule] "it didn't fit the aesthetic of the sub" and I was under the impression that it did. In the future I think I may limit my submissions as I disagree with the sub moderators on what vapor wave aesthetic really is." - P132*

> *"There's always one really subtle rule you don't notice and then it gets removed ." - P707*

These instances of dissatisfaction with community guidelines partly explain the surprising finding from my statistical analyses that users who read the guidelines before posting submissions have negative attitudes about the community moderation (Section 6.4.2).

6.6 Discussion

In this section, I discuss the implications of my findings, focusing on community guidelines, transparency of explanations, and nurturing of dedicated users.

6.6.1 Community Guidelines

In prior research, Kiesler et al. hypothesized that “explicit rules and guidelines increase the ability for community members to know the norms” (Kiesler, Kraut, and Resnick, 2012). My findings add evidence to the value of establishing explicit posting guidelines in online communities. As I show in Section 6.4.2, participants who posted in subreddits containing rules were significantly more likely to consider their removal as fair. However, only about half of all Reddit communities have explicit rules (Fiesler et al., 2018). Since having a list of posting rules is largely a one-time task, site managers should encourage voluntary moderators to establish rules in their communities. Recent research suggests that community norms and rules often overlap among different communities (Chandrasekharan et al., 2018; Fiesler et al., 2018). Moderators of new communities may benefit by having tools that can suggest them which rules they should create, based on the similarity of their community with existing communities containing rules.

I contribute theoretical insights on the role that community guidelines play in shaping user attitudes. In prior research, Kiesler et al. hypothesized and Matias empirically showed that prominently displaying community guidelines helps increase users’ adherence to those guidelines (Kiesler, Kraut, and Resnick, 2012; Matias, 2019). My findings add nuance to this result by showing that simply making users read the community guidelines may be insufficient to improve the long-term health of the community. Indeed, I found that when moderated users read the rules before posting, they are *less* likely to consider their post removal as fair (Section 6.4.2). I bring attention to the attributes of community guidelines that are important to end-users: their size (i.e., number of rules in the guidelines), subjectivity, reason why each rule is created, and effort needed to comply with the guidelines. Next, I discuss how the insights my findings provide about these different attributes of community guidelines may benefit platforms and community managers.

My statistical analysis indicates that when users perceive the community rules to be clear, they are more likely to consider the removal as fair, and are more likely to consider posting again (Section 6.4.2, Table 6.5). In line with this, my qualitative findings suggest that users find it difficult to follow rules that are imprecise, subjective or require a lot of effort to comply with (Section 6.5.5). Thus, when community managers create rules, they should consider whether the rules are clearly laid out and easy to follow.

Prior research has shown that community guidelines are often created as reactions to short-term events or transitions (Seering et al., 2019b). New users, however, may not be aware of these past events. As my analysis shows (Section 6.5.5), some users do not comply with the community rules because they do not understand the reasoning behind why these rules were put in place. Many users do not realize why certain rules are needed. Therefore, documenting the reasons for rule creations or explaining the need for such rules may help increase the acceptability of rules among new users.

When we evaluate the dynamics of users' compliance with community guidelines, it is important to consider that many social media users engage with multiple platforms (Smith and Anderson, 2018). Since different platforms may have different posting guidelines (Pater et al., 2016), it can be challenging for users to precisely follow these guidelines on each platform. My findings suggest that adhering to guidelines becomes even more difficult in a multi-community environment such as Reddit where alongside the site-wide policies, each community has its own set of unique rules. As the list of rules in a community becomes longer, it becomes increasingly onerous for new users to attend to all the rules and make a successful submission. One approach to address this problem could be to introduce a pre-submission step where submitters are asked what type of post they are about to submit. Following this step, only the community rules relevant to that post type may be shown to the submitter. This may reduce the burden of verifying compliance with the entire set of rules for the submitters, and make it easier for them to post successfully.

Platforms can also design to make compliance with certain rules easier on the users. In particular, moderation systems can warn the user when they are about to post a submission that violates a rule whose compliance can be automatically verified. For example, if a community requires posts to be in a certain format, users should be alerted if they have the wrong format at the time of submission, and they should be allowed to edit their post to avoid future removal. This would also help assist users' understanding of the rules and social norms of the community.

6.6.2 Transparency of Explanations

My findings clearly highlight that lack of transparency about content moderation is a key concern among moderated users (Sections 6.4.1, 6.4.3, 6.5.1, 6.5.4). I found that many users felt confused about why their post was removed. Some participants were more frustrated at lack of notification about the removal than about the removal itself. At a broader level, transparency in moderation has important implications for our communication rights and public discourse, as pointed out by prior research (Gill, Redeker, and Gasser, 2015b; Suzor, Van Geelen, and Myers West, 2018).

My survey data show that in absence of information about why removal occurred, users often develop folk theories about how moderation works (Section 6.5.2). While these folk theories may be inaccurate, they influence how users make sense of content moderation and how they behave on the site. Therefore, more examination is needed of how users consume and integrate different information resources to create folk theories about moderation systems. Further, we must update our theories on best practices in moderation systems to account for how users' folk theories about these systems may influence their behaviors.

In their study of how users form folk theories of algorithmic social media feeds, DeVito et al. showed that most folk theories held by the users are "flexible, as opposed to closely-held, rigid beliefs" (DeVito et al., 2018). This suggests that designing "seams" into (Eslami et al., 2016; Jhaver, Karpfen, and Antin, 2018) or providing more accessible information about how moderation systems work may influence users to revise their folk theories and improve their attitudes about the community. Yet, in his study of the effects of system transparency on trust among end-users, Kizilcec found that providing too much information about the systems eroded users' trust (Kizilcec, 2016). Indeed, in recent literature, researchers have begun asking questions about how much transparency is enough (Eslami et al., 2019; Jhaver, Karpfen, and Antin, 2018), and how well transparency in design can serve the outcomes for important values like awareness, correctness, interpretability, and accountability (Rader, Cotter, and Cho, 2018). In Chapter 5, I also show that it is not appropriate for community managers to be fully transparent about how moderation works because bad actors may then game the system and post undesirable content that evades removal (Jhaver et al., 2019a). Therefore, community managers must carefully attend to the design of explanation mechanisms, scrutinizing what and how much information they reveal through these processes. Similarly, they should cautiously consider the pros and cons of notifying the user of post removal versus silently removing posts. I also expect that not all post removals are similar. Future research should explore whether it is advantageous to distinguish between sincere users and bad actors when determining whether to provide removal notifications.

While the possibilities of being exploited by bad actors remains a challenge to implementing transparency in moderation, my results do indicate that informing the users *why* their post was removed through an explanation message can be a useful educational experience. Some of my participants pointed out that they felt more prepared to make successful posts in the future after they received explanations for post removals. My statistical analysis also highlights that when removal explanations provide information that is new to the users, they perceive the removal as more fair (Table 6.6). Therefore, moderators must focus on improving the content of the explanation message, and make it relevant to the moderated posts. Additionally, the mode or source of removal explanation does not seem to matter to the users (Table 6.6). This suggests a potential for building automated moderated tools to deliver explanation messages. Such tools may help improve users' perceptions of content moderation without unduly increasing the work load of human moderators.

6.6.3 Are They Acting in Good Faith?

My qualitative analysis suggests that there is considerable variance among users whose posts get removed. There are those who describe themselves as "trolls" or who post on Reddit for hateful or bigoted reasons (Section 6.5.3). Such users are not invested in contributing valuable information or fostering constructive discussions with others on Reddit communities. These users deliberately post content that they know is likely to get removed.

On the other end of the spectrum, moderated users include those who are emotionally engaged in the social life of Reddit communities. They invest substantial amounts of time and effort in contributing content to share with others on Reddit but mistakenly violate a community guideline or social norm and suffer removal. Such users feel embarrassed or unappreciated when their posts get removed (Section 6.5.1). Some users seek crucial advice from the community in their posts, and they feel dejected when their post is removed, as poignantly expressed by the autistic respondent who sought help from the r/disability community.

Social media platforms like Reddit have opened new avenues through which individuals seek not just informational but also emotional and social support. However, when users get moderated without appropriate feedback, they feel dejected, they are less likely to contribute further, and they may even leave the community. To allow a diverse set of users to participate in the digital public spheres, it is vital that moderation systems support not just the users who are in the know, but also those who may be unaware of the normative practices of online communities.

Just as importantly, I argue that platforms themselves may also benefit from nurturing users who are invested in learning from their mistakes and are just confused about where things went wrong. Users who attempt to post a submission on a community in good faith show a certain amount of dedication to contribute to the community. Therefore, if moderation systems offer them opportunities to participate and grow, they may turn into valuable contributors.

Given the variety of users who post online, moderation systems should find ways to distinguish sincere users from trolls and invest their resources in nurturing the former. Yet, as prior research shows, making such distinctions can pose many challenges (Blackwell et al., 2018c; Blackwell et al., 2018b; Phillips, 2015a). For example, it may be problematic to classify a user as a bad actor if she posts an offensive message in response to another offensive post. Further, we may need to develop different strategies to address different types of bad actors (Blackwell et al., 2018b). It might be worthwhile explaining the posting norms to a new user who posts an offensive joke, but not to another user who repeatedly posts disturbing content.

While it is necessary for moderation systems to deter bad actors, it is also important to nurture sincere users. My statistical analyses show that when users spend more time in creating a post that is subsequently removed, they are less likely to consider posting again (Section 6.4.1). However, online communities need exactly such dedicated users to foster healthy growth. Therefore, information clues that can help moderators identify sincere users can be constructive. For instance, designing tools that allow moderators to easily notice the posts that took a long time for the submitter to create may help the moderators identify and engage with sincere users.

I note that although supporting users who have the potential to be valuable contributors is a worthy goal, there are other constraints and trade-offs that need to be considered.

For example, moderator teams, particularly on platforms like Reddit where voluntary users regulate content, often have limited human resources. Such teams may prioritize removing offensive or violent content to keep their online spaces usable. Additionally, moderators may find more value in consistently enforcing their community guidelines regardless of the motivation of the contributor. Still, my findings suggest that online communities may find it useful to invest their resources towards nurturing users who show dedication to contribute well. For example, even when moderators have to remove sincere posts to consistently enforce community guidelines, they can contact post submitters and provide them actionable feedback on how to post successful submissions in the future.

6.6.4 Limitations and Future Work

Like all online survey studies, this study suffers from self-selection bias and social desirability bias. Users may be biased to obscure essential parts of the moderation experience that may present them in a negative light. Yet, I believe that the subjective perspectives of moderated users I present here provide important insights about the user experience on Reddit and can guide future design and moderation strategies.

While I account for some key demographic factors such as age, education and gender in my analyses, I did not collect data on other important factors like race and sexuality that are likely to influence how users react to content removals. Further, Reddit communities vary among one another on a wide range of factors such as their topic, norms, goals, and policies, but I could not control for such differences in my analyses. Exploring how these factors affect user responses to content moderation could be a productive direction for future work. Another limitation of my sample is that it skews heavily young and heavily male. While I suspect that this may reflect the actual demographics of Reddit users, this still results in an under-representation of the perspectives of other age groups and genders. Future work that focuses on under-represented age groups and gender may provide valuable insights.

I have focused only on post submissions for the purpose of this chapter. However, users also comment on these posts, and moderation of comments may involve a different set of concerns. Analyzing how users perceive comment removals differently from post removals may be a fruitful area for future research.

6.7 Conclusion

I began this study to understand the perspectives and experiences of Reddit users whose posts have been removed. My findings highlight users' frustrations with various aspects of content moderation: absence of removal notifications, lack of explanations about why posts were removed, community guidelines that can be interpreted in multiple ways, and mistaken removals by automated moderation tools, among others. Lack of transparency in moderation resulted in many users creating folk theories about how content moderation happens. Suspicions about the political biases of moderators were commonly held among my participants.

Analyzing my findings led me to reflect on the question: How should we think about "fairness" in the context of content moderation? Is fairness of content removals whatever the community moderators think is appropriate to remove? Is fairness that unpopular users deserve to have their posts removed even if they believed in good faith to have been following the rules?

From the perspectives of my participants, fairness is associated with having a clear set of rules, getting informed when content removal occurs, and receiving explanations for the removal. Yet, is it fair to expect content moderators to invest their limited resources into these tasks? Even if the moderators take on these tasks, it is possible that such investments may not be productive in many instances, and they may open up opportunities for trolls to waste the time of moderators.

Still, it may be worth considering how a greater focus on transparency and explanations may affect the communities. Currently, moderation mechanisms remove content at massive scales, often without notifying the users of removal. On a positive note, my findings show that when moderator teams commit to transparency and provide removal explanations, users sometime learn from their mistakes and they feel better prepared to make successful submissions.

Therefore, an emphasis on education, rather than removal, may improve users' outlook towards the community and encourage them to participate constructively. Designing automated tools that can provide removal explanations in specific scenarios can help newcomers become familiar with the norms of community without unduly increasing the work of moderators. Creating community guidelines that are clear and easy to follow can improve users' perceptions of fairness in moderation. Ultimately, moderated users include many individuals who have made a deliberate effort to contribute to the community. Therefore, nurturing these users and attending to their needs can be an effective way to sustain and improve the health of online spaces.

CHAPTER 7
EVALUATING THE IMPORTANCE OF TRANSPARENCY IN MODERATION: AN ANALYSIS OF USER BEHAVIOR AFTER CONTENT REMOVAL EXPLANATIONS ON REDDIT

7.1 Introduction

Social media platforms like Facebook, Twitter, and Reddit have become enmeshed in a wide range of public activities, including politics (Gillespie, 2017b), journalism (Newman, 2009), civic engagement (Zúñiga, Jung, and Valenzuela, 2012), and cultural production (Nieborg and Poell, 2018). As such, the decisions that these platforms make have a substantial impact on public culture and the social and political lives of their users (Gillespie, 2015; DeNardis and Hackl, 2015). Unfortunately, the black-box nature of content moderation on most platforms means that few good data are available about how these platforms make moderation decisions (Suzor, Van Geelen, and Myers West, 2018). This makes it difficult for end users, particularly those with low technical expertise, to form an accurate mental model of how online content is curated. For example, most of the time on Reddit, content simply disappears without feedback. This lack of transparency of moderation decisions can diminish the comprehensibility of content regulation, which can decrease users' trust in social media platforms.

One strategy for improving transparency is to provide end users with explanations about why their content was removed. Prior research on explanations span a number of different fields such as cognitive science, psychology and philosophy (Herlocker, Konstan, and Riedl, 2000). The importance of explanations in providing system transparency and thereby increasing user acceptance has been demonstrated in many areas: e-commerce environments (Wang and Benbasat, 2007; Pu and Chen, 2006), expert systems (Klein and

Shortliffe, 1994), medical decision support systems (Armengol, Palaudaries, and Plaza, 2001), and data exploration systems (Carenini and Moore, 1998).

What effect does providing explanations have on content moderation? When equipped with the right explanation mechanisms, moderation systems have the potential to improve how users learn to be productive members of online communities. Explanations could provide individualized instructions on how to complete tasks such as making a successful submission or finding the right community for their post. However, this obviously comes with a cost: someone has to spend time crafting and delivering explanations to users whose content has been removed.

In this work, I focus on understanding transparency in content moderation on the popular social media platform Reddit[1]. Reddit has more than a million subcommunities called subreddits, with each subreddit having its own independent content regulation system maintained by volunteer users. In this way, the Reddit platform provides a rich site for studying the diversity of explanations in content management systems and their effects on users.

My analysis is guided by the following research questions:

- RQ1: What types of post removal explanations are typically provided to users?
- RQ2: How does providing explanations affect the future posting activity of users?
- RQ3: How does providing explanations affect the future post removals?

While Chapter 6 explored the effects of removal explanations on the *attitudes* of end-users, the current chapter investigates how providing removal explanations (and the way in which they are provided) is associated with the future *behavior* of users.

I break my analysis into two parts. First, I present a general characterization of removal explanations that are provided on Reddit communities. This characterization provides a descriptive sense of the types of information made available to users whose posts are moderated. Applying topic modeling techniques on a corpus of 22K removal explanations, I

[1] Findings from this study were published in 2019 in the ACM CSCW conference(Jhaver, Bruckman, and Gilbert, 2019). My PhD advisors, Amy Bruckman and Eric Gilbert, guided me on this work.

found that explanations not only provide information about why submissions are removed, they also reveal the mechanics of how moderation decisions are made, and they attempt to mitigate the frustrations resulting from content removals. I also characterize the differences between explanation messages offered through different modes (comments v/s flairs), which I further inspect in my subsequent analyses. Next, I explore quantitative relationships between removal explanations and subsequent user activity. I also analyze how different elements of explanation such as the length of explanation, the mode through which it is provided, and whether it is offered by a human moderator or an automated tool affect user behavior.

My findings show that provision of removal explanations is associated with lower odds of future submissions and future removals. I also find that offering explanations through replying to the submission is more effective at improving user activity than simply tagging the submission with a removal explanation. I build on my findings to provide data-driven guidelines for moderators and community managers in designing moderation strategies that may foster healthy communities. My results also suggest opportunities for moderation systems to incorporate education (over punishment), and I discuss how such a shift may help communities manage content at scale.

This chapter is concerned with the removals of submissions and the subsequent actions taken by the moderation teams. I only focus on moderation of submissions but not comments in this study because my interviews with Reddit moderators in previous studies (Chapter 5, (Jhaver, Vora, and Bruckman, 2017)) led me to believe that explanations for comment removals are provided extremely rarely on Reddit. As I explained in Chapter 6, when a submission is removed on Reddit, moderators can choose to provide the submitter with an explanation for why this removal occurred. This can be done in a variety of ways. For example, moderators can comment on the removed post with a message that describes the reason for removal (Figure 7.2). Alternatively, they can flair the removed post (Figure 7.3) or send a private message to the submitter. Moderators can either choose to compose the removal explanation themselves, or they can configure automated tools (e.g., AutoModerator (Chapter 5)) to provide such explanations when the submission violates a community guideline.

My analysis focuses on how content removals affect future user behaviors on Reddit. I quantify the user behavior using two measures: (1) whether the user posts a submission, and (2) whether the user's posted submission gets removed. I also explore how providing explanations and the different attributes of explanations affect these measures of user behavior.

In the rest of this chapter, I begin by describing how I collected and prepared data to answer my research questions for this study. I then present an overview of explanations for content removals that are provided on Reddit communities using topic modeling and n-gram analyses. Next, I describe my methods, before detailing my findings on how explanations are associated with future activity of Reddit users. Finally, I use the insights gained from my overview of removal explanations to ground my quantitative results, and articulate the lessons learned from my research for the benefit of site managers, moderators, and designers of moderation systems.

7.2 Data Preparation

I first collected a dataset D of all (allowed as well as removed) Reddit submissions that were posted over the eight months period March 2018 to October 2018. I downloaded this data using the pushshift.io service[2]. As I mentioned above, I only focus on moderation of submissions but not comments in my analyses. Following the ethical recommendations from prior research (Chancellor, Lin, and De Choudhury, 2016), I did not collect submissions that were deleted by their posters in this data. This dataset contained 79.92 million submissions, out of which 17.40 million submissions (21.77%) submissions were removed.

I wanted to explore how moderation decisions and removal explanations on prior posts

[2] https://files.pushshift.io/reddit/submissions/

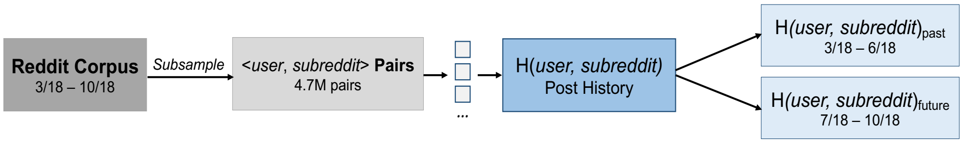


Figure 7.1: Flowchart depicting the data preparation. I collected posting history for a sample of <*user, subreddit*> pairs between March and October 2018. Next, I split this posting history for each pair and aggregated posts to create H_{past} and H_{future} datasets.

of a user in a Reddit community affects the future posting behavior of that user in that community. For this, I began with identifying and removing the submissions made by bots in my data. First, I identified Reddit bot accounts by collecting a list of known bot accounts on Reddit (RedditBots, 2019) which included "AutoModerator." Analyzing the patterns of bot user-names on this list, I also considered accounts whose user-names ended with "Bot", "_bot", "–bot" or "Modbot" to be bot accounts. I also manually reviewed the user profile and posting history of accounts that posted more than 10,000 submissions, and identified accounts that were clearly bots. As there is no way to be fully certain whether a given Reddit account is a human or bot account, I acknowledge that my method only approximates distinguishing between human and bot accounts.

After identifying the bot accounts, I removed all the submissions posted by these accounts from my original dataset *D*. Next, I sampled a set of 4,705,048 <user, subreddit> pairs by retrieving all the unique <u, s> pairs where user u posted a submission s in the month of July. Following this, for each <u, s> pair, I retrieved the entire posting history H(u, s) of u in s between the period March 2018 and October 2018 (Figure 7.1). In total, this data sample, *S*, consisted of 32,331,120 submissions.

I split the posting history H(u, s) for each <u, s> pair in two groups - $H(u, s)_{\text{past}}$ and $H(u, s)_{\text{future}}$. The $H(u, s)_{\text{past}}$ group contains all submissions prior to and including the first submission made by u in s since the start of July 1, 2018 (mid-point of my dataset), and the $H(u, s)_{\text{future}}$ group contains all the remaining submissions made by u in s. I aggregated all submissions in $H(u, s)_{\text{past}}$ and $H(u, s)_{\text{future}}$ into the datasets H_{past} and H_{future} respectively.

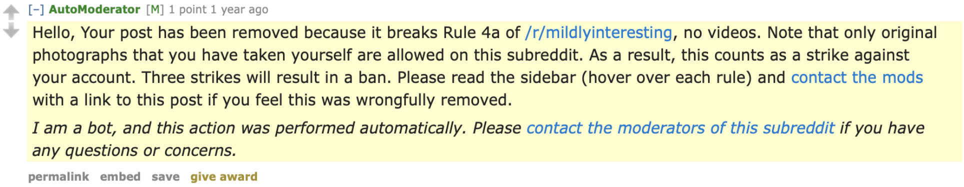
7.2.1 Collecting Removal Explanations

To analyze the effects of past removal explanations on future behaviors, I next collected the removal explanations for removed posts in H_{past}. For each removed post in H_{past}, I first collected all the comments posted as direct replies to that post as well as the flairs assigned to the post. To distinguish removal explanation comments from other comments, I first examined all comments to a random sample of 500 removed submissions. I manually identified the comments that provided a removal explanation and were authored by one of the moderators of the corresponding subreddit or an automated moderation bot. Through inspection of these comments, I obtained an initial *removal phrases list* of 24 phrases that frequently occurred in removal explanation comments but not in other comments. These phrases included “doesn’t follow rule”, “submission has been removed” and “feel free to repost.”

Following this, based on a snowball sampling approach, I filtered all the comments to the removed submissions in H_{past} containing any of these seed phrases and added more phrases incrementally through manual inspection of those comments. With addition of each phrase to *removal phrases list*, I retrieved a sample of comments to removed submissions containing only that phrase and verified that the obtained comments were removal explanations. This process expanded my *removal phrases list* to 93 phrases. I searched for comments containing any of these phrases to retrieve a list of removal explanation comments.

I adopted a similar approach to create a *removal phrases list* for flairs that contained 32 phrases which included “removed”, “karma” and “low effort.” I used this list to distinguish removal explanation flairs from other flairs. I also distinguished removal explanation comments that were provided by bot accounts from those authored by human moderators. I identified bot accounts for this step using the same approach as described at the beginning of this section.

This process resulted in a collection of 212,748 explanation messages. To evaluate the



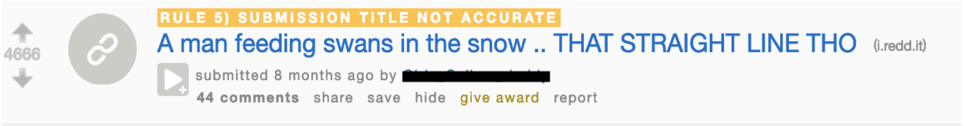
[–] AutoModerator [M] 1 point 1 year ago

Hello, Your post has been removed because it breaks Rule 4a of /r/mildlyinteresting, no videos. Note that only original photographs that you have taken yourself are allowed on this subreddit. As a result, this counts as a strike against your account. Three strikes will result in a ban. Please read the sidebar (hover over each rule) and contact the mods with a link to this post if you feel this was wrongfully removed.

I am a bot, and this action was performed automatically. Please contact the moderators of this subreddit if you have any questions or concerns.

permalink embed save give award

Figure 7.2: An example explanation message provided through a comment to a removed submission.



4666

RULE 5) SUBMISSION TITLE NOT ACCURATE

A man feeding swans in the snow .. THAT STRAIGHT LINE THO (i.redd.it)

submitted 8 months ago by [redacted]

44 comments share save hide give award report

Figure 7.3: An example explanation message provided by flairing the removed submission. Username has been scrubbed to preserve the anonymity of the submitter.

data quality, I randomly sampled 200 of these messages and manually reviewed them to see whether they provided information about post removals. This analysis found only 3 messages that were not explanations for post removals, which indicates that my approach to identify explanation messages has a high precision.

I note that some moderators may have provided removal explanations through private messages to post submitters. Because I did not have access to private messages, my data is missing these explanations. Therefore, my results should be interpreted taking this limitation into account.

7.3 An Overview of Removal Explanations

I begin by characterizing the removal explanations in my dataset and then provide an overview of how explanations provided through different modes differ from one another.

As described in Section 7.2, I extracted all the messages that explained why submissions were removed in H_{past}. Figure 7.2 shows an example of removal explanation posted by AutoModerator, an automated moderation bot popular on Reddit, as a reply comment to the removed submission. This message describes in detail the specific rule the submitter seems to have violated, the negative consequences of future rule violations, and the steps the user can take to appeal against the moderation decision. Figure 7.3 shows a removed submission that has been flaired with the message: “Rule 5) Submission Title not Accurate.” This message specifies the community rule the submitter has broken, but doesn’t provide any other contextual information.

Overall, I found that 207,689 removed submissions in H_{past} were provided removal explanations. 10.72% (N = 22,269) of these submissions received explanations only through a comment, and 86.84% (N = 180,357) received explanations only through a flair. 2.44% of removed submissions (N =5,059) received explanations through a comment as well as a flair. This shows that Reddit communities use flairs much more frequently than comments as a mechanism to present the reasoning behind their moderation decisions. The average length of removal explanations provided through comments was 728.81 characters (median = 572, SD = 510.12), whereas the average length of flair explanations was 17.34 characters (median = 16, SD = 10.44). This indicates that explanations offered through comments are usually much more detailed than explanations provided using flairs.

Next, I sought to separate explanations provided by human moderators from those provided by automated tools. Because the Reddit API does not provide information on which Reddit accounts are responsible for flairing any submission, I could not calculate how many of the explanation flairs were produced by automated tools and how many were produced by human moderators. However, explanations provided through comments contained information about which Reddit account posted them (e.g., see Figure 7.2). Analyzing these explanations, I found that 58.18% of all comment explanations were provided by known bots. This shows that automated tools have come to play a critical role in moderation systems not just for removing inappropriate content but also for associated tasks such as explaining moderation decisions. I also found that a great majority (94.62%) of these automated explanations were provided by “AutoModerator,” a moderation tool offered to all subreddits (Chapter 5).

Next, I analyzed the removal explanation messages in order to get a descriptive understanding of the content of these messages. I began by applying standard text-processing steps such as lowercasing all characters, removing special characters and excluding stop words from comments and flairs. I also discarded all hyperlinks that appeared in these data.

I adopted the Latent Dirichlet Allocation (LDA) technique (Blei, Ng, and Jordan, 2003) to extract the range of different types of explanations provided through comments. LDA is a widely used statistical model to discover latent topics in a collection of documents in which each topic consists of a set of keywords that defines it, and text tokens are distributed over latent topics throughout each document. I treated each explanation as a document and applied LDA on all comment explanations. I chose the number of topics, *k*, for this model based on perplexity scores (Wallach et al., 2009), a useful measure for comparing and selecting models and adjusting parameters (Newman et al., 2010). Testing for different values of *k*, I found that the perplexity score for the model dropped significantly when *k* increased from 5 to 28, but did not change much from 28 to 50. I also looked at the topics themselves and the highest probability words associated with each topic when using different values of *k* to consider if the structure made sense. Through this process, I determined to use 28 topics for comment explanations.

Table 7.1 lists the top ten topics for explanations provided through comments. I manually labeled the topics and measured the relative frequency with which each topic occurs in the data. Specifically, given $\theta(topic_i) = \Sigma p(topic_i|comment)$ over all comments, the values in the parentheses in the first column correspond to $\theta(topic_i)/\Sigma_j\theta(topic_j)$, expressed as a percentage. This table also shows the corresponding keywords as well as explanation examples from each topic.

As can be observed in Table 7.1, explanations offered through comments often provide information about the reason why the submitter's post was removed. For example, topics like "Low karma" and "Flair posts before submitting" suggest attempts to explain to the users why their post warranted expulsion. However, I also found topics like "Removal is

Table 7.1: Top Topics from LDA model of comment explanations with snippets of example messages. All user names in examples have been converted to ‘username’ to preserve the anonymity of users.

Lexical Group and Topic Terms	Examples
Removal is automatic (6.72%): “automatically”, “compose”, “performed”, “contact”, “bot”, “action”, “concerns”	“Your submission has been automatically removed.*I am a bot, and this action was performed automatically. Please contact the moderators of this subreddit if you have any questions or concerns.*”
Low karma (6.48%): *karma, threshold, spammers, banned, subreddits, note, automatically*	“Hello /u/username! Unfortunately, your post was automatically removed because you do not exceed my karma threshold. This has nothing to do with rule violations, it just means that your account is either too new, or doesn’t have enough karma. I have a threshold to prevent spammers from posting on /r/dankmemes.”
Flair post before submitting (6.18%): *“flair”, “science”, “forum”, “post”, “medical”, “wait”, “questions”*	“Hi username thank you for submitting to /r/Askscience.**If your post is not flaired it will not be reviewed.** Please add flair to your post. Your post will be removed permanently if flair is not added within one hour. You can flair this post by replying to this message with your flair choice.
Ask questions in post title (5.64%): “question”, “title”, “answers”, “post”, “please”, “edited”, “mark”	“Your post has been removed as it violated [Rule 1] because it did not end with a question mark.* You must post a clear and direct question, **and only the question**, in your title. * Do not include answers or examples in the post title. You can post answers as comment replies when you’ve reposted.* Please combine clarifying sentences into the question itself.”
Title must describe content (5.12%): “content”, “original”, “allowed”, “outline”, “esque”, “indicating”, “opening”	“Hi username, thank you for posting on /r/oddlysatisfying. Unfortunately, your post has been removed for the following reason:* **Rule 5)** The title of the submission must describe the content it shows.”
Removal is unfortunate (4.93%): *submission, unfortunately, removal, contact, action, concerns, please, questions*	“Thank you for your submission! Unfortunately, your submission has been automatically removed because it contains the phrase ELI5, so it is possible you are looking for /r/explainlikeimfive. ”
Don’t post easily searchable questions (4.85%): “thread”, “easily”, “question”, “questions”, “daily”, “topic”, “reach”	"“Hey there, /u/username! Thanks for your submission, but unfortunately we’ve had to remove your post as it doesn’t follow Rule 3 - No limited scope or easily searchable questions. These types of question belongs in my [Daily Question Thread] and are not allowed as standalone posts”
Check rules in the sidebar (4.24%): “sidebar”, “check”, “thinking”, “appreciate”, “search”, “quite”, “rules”	“Hey there, friendo u/...! Thanks for submitting to r/wholesomememes. I loved your submission, *r/askquija can be wholesome*, but it has been removed because it doesn’t quite abide by my rules, which are located in the sidebar.”
Rule number that has been violated (4.11%): “rule”, “removed”, “violating”, “breaking”, “thank”, “following”, “months”	“Removed, rule 1.”; “Removed for rule 5”
Submission must be a direct image link (3.93%): *imgur, jpg, gif, png, links, albums, longer*	“I are no longer accepting imgur albums as submissions. Please re-submit each individual picture in the album using a direct image link (must end in jpg, gif, png, etc). Thanks. [These instructions might help.](http://i.imgur.com/RjrqakK.gifv) ”

automatic” and “Submission must be a direct image link” which suggest efforts by moderators to make the process of automated moderation and its limitations more explicit to the users. Topics such as “Removal is unfortunate” indicate an effort to gain the confidence of the users and to cushion against the dissatisfaction resulting from the removal. I also found topics on normative guidelines such as “Check rules in the sidebar” that go beyond just the specific post in question and educate users on how to become more central members of the community.

I did not apply LDA on explanations provided through flairs because flair explanations were often too short (median length = 16 characters) to obtain valid insights using LDA modeling. In lieu of this, I extracted unique unigrams, bigrams and trigrams from removal explanation flairs and counted their frequencies in the corpus of flair explanations. n-gram refers to a contiguous sequence of n words from text. Table 7.2 list the most frequent unigrams, bigrams and trigrams for flairs. This table suggests that explanations provided through flairs do not seem to employ hedging phrases as frequently as comment explanations. They appear to be much more direct and to the point, with many common phrases like “non whitelisted domain”, “overposted content” and “r3 repost removed” referring to the subreddit rule the submitter seems to have broken.

I will build upon the differences between comment and flair explanations identified in this section to analyze later in Section 7.5 whether these differences are associated with variations in future activity of moderated users.

7.4 Relationship with Future Activity

Building on the descriptive characteristics of explanations in the previous section, I turn to the relationship between explanations and relevant user activity measures. In this section, I describe how I developed my analytic models on *S*, the dataset containing the posting history of my 4.7 million <user, subreddit> pairs, to answer my research questions. I also discuss the simplifying assumptions I made for these analyses.

Table 7.2: Frequent phrases from Removal Explanation Flairs

Unigram		Bigram		Trigram	
Phrases	**Frequency**	**Phrases**	**Frequency**	**Phrases**	**Frequency**
removed	65817	removed rule	20149	non whitelisted domain	2671
rule	53902	low karma	7038	rule overposted content	2252
fluff	24773	removed repost	3119	rule non gore	1350
repost	17485	fluff question	2896	r14 social media	1179
low	10737	non whitelisted	2671	social media sms	1179
submitted	10238	whitelisted domain	2671	media sms removed	1179
karma	7138	low effort	2593	removed crappy design	1151
title	6816	repost removed	2548	use approved host	1110
content	5406	rule overposted	2252	approved host removed	1110
post	4397	overposted content	2252	assign flair post	979
non	4299	appropriate subreddit	1629	low effort meme	902
shitpost	3812	rule repost	1602	removed restricted content	875
question	3774	social media	1594	removed location missing	849
domain	3387	rule animeme	1528	removed low quality	715
r1	3254	rule non	1491	r3 repost removed	637

I applied logistic regression analyses on *S* for their ease of interpretability after checking for the underlying assumptions. I built these models in such a way that the independent variables derive from characteristics of submissions in the H_{past} group and the dependent variables derive from information about submissions in the H_{future} group. In this way, I am able to analyze the relationship between moderation actions on past submissions and future user activity.

My aim was to use these statistical models to investigate the different aspects of removals and explanations and present results on how they relate to future user submissions and content removals. By splitting the post history H(u, s) for each <u, s> pair into H(u, s)$_{\text{past}}$ and H(u, s)$_{\text{future}}$ at the same time, my analyses aimed to control for the external events and temporal factors that may have affected future user behaviors across different <user, subreddit> pairs. For removed submissions that received explanation through a comment as well as a flair, I chose to ignore the flair explanation and considered only the comment explanation as comments are usually much longer and more informative than flairs. I do not make causal claims that the moderation practices I explore in this study lead to improved user behavior. Rather, I am providing evidence that explanations play some role in determining the users' future activity on Reddit communities.

I note that many Reddit users post on multiple subreddits, and moderation actions in one subreddit may affect user behavior in other subreddits in the future. For example, if a user's submission on the r/science subreddit is removed with the explanation message asking that user to read the subreddit rules before posting, this incident is likely to influence the user to read the community rules when posting on any other subreddit too. However, I make a simplifying assumption of treating different <user, subreddit> pairs for the same user as statistically independent in my analyses.

I explored in a separate analysis how filtering the dataset further to include only the subreddits that are active[3] would affect my results and found that the regression analyses on this filtered dataset produced very similar results. Therefore, I only present my results on the dataset without the additional filter for active subreddits. Next, I list the variables that I use in my analyses for each <user u, subreddit s> pair.

7.4.1 Dependent Variables

My descriptive analyses of the data showed that the future number of submissions had a mean of 3.25 and a median of 0. I also found that the future number of removals across my dataset had a mean of 1.77 and a median of 0. Since median is a robust statistical measure of the data, I chose to focus my analyses on whether the future submissions and removals exceed their median value of 0:

1. **Future Submission**: This is a binary variable that indicates for each <u, s> pair whether the user u has a submission in H(u, s)$_{\text{future}}$.

2. **Future Removal**: This binary variable indicates for each <u, s> pair whether the user u has a submission in H(u, s)$_{\text{future}}$ that was removed.

[3] For these analyses, I considered subreddits that received more than one submission per day on average over the four months period March - June 2018 as active subreddits.

7.4.2 Control Variables

Subreddit Variables

For each subreddit s in my sample, I measured these subreddit features in the month of July, the midpoint of my dataset:

1. **Subreddit Subscribers**: Number of subscribers in subreddit s .
2. **Subreddit Submissions**: Total number of submissions posted in subreddit s.
3. **Net Subreddit Removal Rate**: Percentage of all submissions posted in subreddit s that were removed.

These variables are indicative of the size, activity level, and moderation rate of each Reddit community. I control for these variables because I suspected that they are likely to have an effect on user activity. I note that Reddit communities differ among one another on many other important variables which are likely to have an impact on user behaviors. For example, subreddits have different community guidelines, behavioral norms (Chandrasekharan et al., 2017b; Chandrasekharan et al., 2018), topics, rates of user activity, and age, among other factors, all of which are likely to influence user responses to content moderation. Since I do not account for these variations in my large-scale analyses, my statistical models are simplifications of the community dynamics on Reddit.

Post History Variables

These variables measure the number of submissions made by user u in subreddit s and the average community response measured through the number of upvotes and number of comments received by those submissions. Distributed moderation attained through community response is critical to determining how prominently each post appears on Reddit (Lampe and Resnick, 2004). Therefore, I suspected that these variables are likely to have an effect on future user activity. I distinguished community response from centralized moderation actions because I wanted to focus on the role of explicit regulation decisions made by the moderation team. Post history variables include:

1. **Past Submissions**: Number of submissions in H(u, s)$_{\text{past}}$.
2. **Average Past Score**: Average score (determined by the number of upvotes and downvotes) received by the submissions in H(u, s)$_{\text{past}}$.
3. **Average Past Comments**: Average number of comments received by the submissions in H(u, s)$_{\text{past}}$.

I note that although these variables capture some basic features of community responses, I do not account for the nuances of feedback in user-to-user messages. Such feedback may also affect user attitudes about future postings.

7.4.3 Independent Variables

I operationalized a set of independent variables to capture different aspects of content moderation and measure their impact on users. I discuss these variables below:

1. **Past Removal Rate**: Percentage of submissions in H(u, s)$_{\text{past}}$ that were removed. Intuitively, as the proportions of post removals in a community increase, users are less likely to post in the future. I also suspect that with increasing removals, users may learn from their mistakes and are less likely to post submissions that will be removed. Therefore, I predict that past removal rate will have a negative association with both whether the user submits posts in the future and whether the submitted posts are removed.
2. **Explanation Rate**: Percentage of removed submissions in H(u, s)$_{\text{past}}$ that were provided a removal explanation. My hypothesis is that if users receive explanations for a greater proportion of their removed submissions, it can provide them an understanding of the ways in which they falter in their postings, and help them become more

productive in the future. I expect that explanations, as opposed to silent removals, indicate to the moderated users that the community is dedicated to providing transparency in its regulation, and the moderators are willing to engage and work with them. Thus, I predict that explanation rate will be associated with both future postings and future removals. I note that this variable is defined only for <u, s> pairs where user u had at least one removed post in H(u, s)$_{\text{past}}$.

3. **Average Explanation Length**: Average length of all explanations offered to user u in H(u, s)$_{\text{past}}$. I expect that longer explanations are likely to be more comprehensive and provide more details to the moderated user on why they were moderated and what steps the user can take to better attend to the social norms of the community. Thus, I hypothesize that an increase in explanation length will be linked to the future activity of users. I measured this length by calculating the number of characters in the explanation messages. This variable has meaningful values only for <u, s> pairs where user u had at least one removed post in H(u, s)$_{\text{past}}$ that was provided an explanation.

4. **Explanation through Comments Rate**: Percentage of removal explanations that were provided through a comment to the removed submission. Section 7.3 highlights some of the differences between the explanations provided through these two modes. I use this measure to test whether providing explanations through comments as opposed to using only a flair has a significant relationship with the future activity. I note that this variable is defined only for <u, s> pairs where user u had at least one removed post in H(u, s)$_{\text{past}}$ that was provided an explanation.

5. **Explanation by Bot Rate**: Percentage of removal explanations provided by Reddit bots. I expect that when human moderators provide an explanation as opposed to a bot, the explanations are likely to be more accurate and specific to the context of the post. I also suppose that users are likely to appreciate the individualized attention of

human moderators more than an automatic message by a bot. It is also possible that users may consider explanation messages more seriously if they are reprimanded by a real person instead of a bot. Therefore, I hypothesize that a decrease in this rate or a corresponding increase in the rate of explanations provided by human moderators will be linked to an increase in the future activity of users and reduce instances of post removals. Note that Reddit API does not provide any information on which user account flaired a post. Therefore, I have calculated this rate only for explanations provided through comments. As a result, this variable has meaningful values only for <u, s> pairs where user u had at least one removed post in H(u, s)$_{\text{past}}$ that was provided an explanation through a comment.

I note that although the independent variables discussed above capture many important aspects of Reddit moderation that may affect user behavior, there are other factors that I do not control for in my analyses. For example, I could not account for how moderated users may be affected by their private conversations with moderators in cases where they appeal to reverse the moderation decisions because I do not have access to these conversations. Further, I could not control for how users' demographic characteristics such as their gender, race, age, and education affect their responses to content moderation. Therefore, I see my models as reasonable simplifications of the complex sociotechnical system of Reddit.

It should also be noted that community managers may provide removal explanations for reasons that go beyond providing transparency about moderation decisions. For example, this may be a signaling mechanism for the moderators to communicate to the community members that the subreddit is actively monitored, or this may indicate to the fellow moderators that a specific post has already been reviewed. Regardless of the specific motivations that drive different moderators to provide explanation messages, these messages provide users greater insight into the moderation processes, and my analyses seek to explore the effects of these messages on user behaviors.

7.5 Findings

In this section, I use logistic regression models to examine the influence of independent variables on the dependent variables identified in the last section.

7.5.1 Descriptive Statistics

To begin, I report in Table 7.3 the descriptive statistics for all the variables I have introduced in the previous section, before entering into the regression models. As I mentioned in section 7.4.3, some of these variables do not have any meaningful value in many instances because of the way that they are defined. For example, "Explanation Rate" does not have any meaningful value for <u, s> pairs where user u did not have any removed posts in H(u, s)$_{\text{past}}$. Thus, I create separate models for evaluating different sets of independent variables with each model containing only the valid entries for the variables considered. Table 7.3 lists the number of valid entries for each variable.

I found that across all (u, s) pairs, users posted an average of 3.62 submissions (median = 1) in the corresponding subreddit in H(u, s)$_{\text{past}}$. Past submissions received a median score of 3.5 (mean = 100.81) and a median of 3 (mean = 10.19) comments. My analysis shows that in 37.5% of all cases (N = 1.73M), user u had at least one future submission in subreddit s. I also saw that for instances where users posted on the corresponding subreddit in the future, a future removal occurred in 31.2% (N = 550.5K) of the cases. The median number of subreddit subscribers is 91.5K and the median net count of all subreddit posts is 1,471. This suggests that a majority of the users submit posts in large, active subreddits. Past submissions were posted in subreddits that removed a median of 14.82% of all submissions.

I created four regression models for each of the two dependent variables. I began creating each new model by first discarding all the cases with any missing data for the variables in the model. This was done to analyze the role of the additional variables in each subsequent model by focusing only on the cases where the variable value is meaningful. Table

Table 7.3: Descriptive statistics (min, max, mean, median, frequency distribution, number of valid entries) of all introduced variables. The distributions of post history and independent variables are shown at a logarithmic scale on y axis.

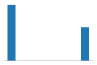
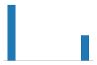
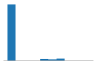
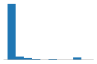
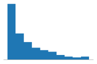
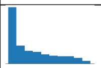
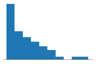
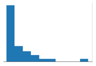
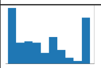
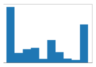
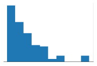
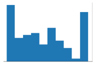
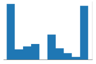
Variable Group	Variable Name	Min	Max	Mean	Median	Distribution	Valid entries
Dependent variables	Future Submission (binary)	0	1	0.37	0		4.7M (100%)
	Future Removal (binary)	0	1	0.31	0		1.8M (37.5%)
Subreddit variables	Subreddit Subscribers	0	31.8M	1.7M	91.5K		4.7M (100%)
	Subreddit Submissions	1	187.5K	12.7K	1,471		4.7M (100%)
	Net Subreddit Removal Rate	0	100	23.38	14.82		4.7M (100%)
Post history variables	Past Submissions	1	19.71K	3.62	1		4.7M (100%)
	Average Past Score	0	266.2K	100.8	3.5		4.7M (100%)
	Average Past Comments	0	72K	10.19	3		4.7M (100%)
Independent variables	Past Removal Rate	0	1	0.25	0		4.7M (100%)
	Explanation Rate	0	1	0.09	0		1.4M (29.3%)
	Average Explanation Length	2	9.9K	152.4	20		147.8K (3.1%)
	Explanation through Comments Rate	0	1	0.20	0		147.8K (3.1%)
	Explanation by Bot Rate	0	1	0.41	0		31K (0.7%)

Table 7.4: Descriptions of statistical models used in my analyses. For each model, the input and output variables, criterion for including data, and the number of valid data entries are shown.

Output Variable	**Model**	**Input variables**	**Inclusion criteria**	**Valid entries**
Future Submission	A.1	Subreddit variables + Post history variables + Past Removal Rate	All <user, subreddit> pairs	4.7M
	A.2	+ Explanation Rate	Past Removal Rate > 0	1.4M
	A.3	+ Average Explanation Length + Explanation through Comments Rate	Explanation Rate > 0	147.8K
	A.4	+ Explanation by Bot Rate	Explanation through Comments Rate > 0	31K
Future Removal	B.1	Subreddit variables + Post history variables + Past Removal Rate	Future Submissions > 0	1.8M
	B.2	+ Explanation Rate	Future Submissions > 0 AND Past Removal Rate > 0	548.7K
	B.3	+ Average Explanation Length + Explanation through Comments Rate	Future Submissions > 0 AND Explanation Rate > 0	64.8K
	B.4	+ Explanation by Bot Rate	Future Submissions > 0 AND Explanation through Comments Rate > 0	15.2K

7.4 describes what variables and data are included in each model and the number of data points for that model. Tables 7.5 and 7.6 show the results of these regression models.

For ease of comparing the relative importance of the explanatory variables, I standardized all the predictor variables in my models so that each variable had a mean of zero and a standard deviation of one. I report the results of my analyses as odds ratios (OR), the change in the odds of posting a submission or experiencing a removal in the future when an input variable is increased by one standard deviation. Odds ratios greater than one indicate an increase in the odds of the corresponding dependent variable, while odds ratios less than one indicate a decrease in the odds. For each model, I verified that multicollinearity was not a major problem as none of the correlations were higher than 0.5. I note that direct comparisons between the Nagelkerke R Square of different models in Tables 7.5 and 7.6 are not possible as each model is composed of a separate subset of the entire data.

7.5.2 Future Submission

In this section, I discuss the results of several regression models for the dependent variable, "Future Submission," a binary variable that indicates for each <u, s> pair whether the user u has a submission in H(u, s)$_{\text{future}}$ (see Tables 7.4 and 7.5).

Observation 1: High past removal rate for the user is associated with lower odds of posting in the future.

I first created a model A.1 using all the control variables (the subreddit variables as well as the post history variables) and past removal rate (Table 7.4). Model A.1 reports the main effects of the control variables and past removal rate on future submissions. It shows that past number of submissions overwhelmingly determines (OR = 7.4E+10) whether the user posts in the future. This is as we would expect—people who are in the habit of posting often will probably continue to post. Beyond this, since the odds ratio for past removal rate is 0.978, one standard deviation (SD) increase in the past removal rate for u in s was associated with 2.2% (100 * (1 - .978)) lower odds of u posting in the future. Intuitively, when

Table 7.5: Odds ratio of predicting whether the user will post in the future. Here, p<0.001: ***; p<0.01:**; p<0.05:*. Each model was constructed on the corpus for which all the included variables have valid entries. The results show that higher *past removal rate* and higher *explanation rate* are associated with a decrease in the odds of users posting in the future. In contrast, higher *explanations through comments rate* and higher *explanations by bot rate* are linked to an increase in the odds of users posting in the future. Other variables are used as controls.

Group	Variables	Model A.1	Model A.2	Model A.3	Model A.4
Subreddit variables	Subreddit Subscribers	0.968***	0.988***	0.984*	0.981
	Subreddit Submissions	1.118***	1.164***	1.146***	1.151***
	Net Subreddit Removal Rate	0.940***	0.938***	0.900***	0.86***
Post history variables	Past Submissions	7.4E+10***	4.5E+6***	687.6***	16.628***
	Average Past Score	0.995 ***	0.999	0.988	0.972
	Average Past Comments	1.037 ***	1.014**	1.043***	1.057**
Independent variables	Past Removal Rate	0.978 ***	0.638***	0.563***	0.520***
	Explanation Rate		0.988***	0.686***	0.636***
	Average Explanation Length			1.003	0.990
	Explanation through Comments Rate			1.035***	0.824***
	Explanation by Bot Rate				1.264***
	# Obs	4.7M	1.4M	147.8K	31K
	Intercept	1.135***	1.154***	1.218***	1.355***
	Nagelkerke R Square	0.191	0.267	0.337	0.327
	Omnibus Tests of Multiple Coefficients	p <.001	p <.001	p <.001	p <.001

users' posts continue to get removed on a subreddit, they may feel that their contributions are not welcome and stop posting, or in some cases, even leave the subreddit.

The odds ratio for the net subreddit removal rate is 0.94. This suggests that an overly strict moderation policy may have a chilling effect on users and inhibit their future postings. I also found that the odds that user u posts in subreddit s in the future increases by 11.8% (100 * (1.118 - 1)) with each standard deviation increase in the net number of submissions that s receives. This shows that regardless of other factors, users are likely to continue posting in active communities. My results also indicate that community engagement with the user posts has a positive effect on future submissions. For example, since the odds ratio for past comments is 1.037, users who received one standard deviation increase in comments on their past submissions are 3.7% more likely to post in the future. Surprisingly, the odds of future posting reduced with increase in the number of subreddit subscribers (OR = 0.968). The average past score had a much smaller effect on future submissions (OR = 0.995).

Observation 2: Greater explanation rates characterize reduced odds of posting in the future.

Next, I created model A.2 to test the relationships between provisions of explanations and the occurrence of future submissions. This model makes a simplifying assumption that the users who received the explanation messages noticed and read them. I only considered cases where the user u had at least one post removal in the past to build this model. I found that explanation rate adds significantly to the model even after controlling for subreddit characteristics, post history variables and past removal rate. Since the odds ratio is 0.988, one standard deviation increase in explanation rate was associated with 1.2% decrease in the odds of future submissions. One explanation for this association is that receiving removal reason messages makes users realize that their posts are being carefully reviewed, and this may make users become more cautious in their posting behavior.

I note that this result has different implications for different types of communities. For example, consider a small community that receives 100 posts a month. Assuming that the relationship between explanations rate and future posts in model A.2 applies to this community, if explanations rate is increased by one standard deviations, this community may have 1.2% fewer posts or about 99 posts a month in the future. In contrast, the same increase in explanations rate would cause a large community that usually receives 10,000 posts a month to have 120 fewer posts a month in the future. Thus, communities must consider how much decrease in traffic they can withstand when determining whether to provide explanations.

Observation 3: Having a higher fraction of explanations offered through comments, rather than through flairs, is associated with an increased likelihood of users posting in the future.

Following this, I built model A.3 to evaluate how different attributes of removal explanations affect user behavior. I only used cases where the user u received at least one removal explanation for his or her past removal to build this model. My results in Table 7.5 show that explanation length did not add significantly to the model for the occurrence of future submissions (OR = 1.003). Thus, my hypothesis that longer explanations are more comprehensive and are therefore more likely to influence greater user engagement was not supported. This model, however, showed that given a fixed number of explanations, providing explanations through comments rather than through flairs is likely to cause an increase in the occurrence of future submissions (OR = 1.035).

Observation 4: Explanations provided by human moderators, rather than by automated tools, are associated with lower odds of moderated users posting in the future.

Finally, I created model A.4 to test the effects of sources of removal explanations. I only used instances where users were provided at least one explanation through a comment to the removed submission to build this model. Because the odds ratio for explanations by bot rate is 1.264 (Table 7.5), this model showed that one standard deviation increase in the rate of explanations provided by bots was associated with 26.4% increase in the occurrence

Table 7.6: Odds ratio of predicting whether the user's post will get removed in the future. Here, p<0.001: ***; p<0.01:**; p<0.05:*. Each model was constructed on the corpus for which all the included variables have valid entries. The results show that higher *past removal rate* is associated with an increase in the odds of user experiencing a post removal in the future. In contrast, higher *explanation rate* and higher *explanations through comments rate* are linked to a decrease in the odds of user experiencing a post removal in the future. Other variables are used as controls.

Group	Variables	Model B.1	Model B.2	Model B.3	Model B.4
Subreddit variables	Subreddit Subscribers	0.910***	0.934***	0.891***	0.93**
	Subreddit Submissions	1.168***	1.033***	1.11***	1.079**
	Net Subreddit Removal Rate	2.461***	2.215***	2.058***	2.021***
Post history variables	Past Submissions	1.164***	2.6***	5.636***	2.443***
	Average Past Score	0.991***	0.981***	0.96**	0.975
	Average Past Comments	1.02***	1.000	1.027	1.0
Independent variables	Past Removal Rate	1.968***	1.366***	1.236***	1.286***
	Explanation Rate		0.935***	0.701***	0.649***
	Average Explanation Length			1.003	1.002
	Explanation through Comments Rate			0.905***	0.774***
	Explanation by Bot Rate				1.019
	# Obs	1.8M	548.7K	64.8K	15.2K
	Intercept	0.392***	2.148***	2.287***	2.044***
	Nagelkerke R Square	0.378	0.187	0.199	0.231
	Omnibus Tests of Multiple Coefficients	p <.001	p <.001	p <.001	p <.001

of future submissions. Equivalently, explanations provided through human moderators are linked to reduced odds of user submitting posts in the future.

7.5.3 Future Removals

In this section, I analyze which factors are associated with whether a post removal occurred for submissions made in H(u, s)$_{\text{future}}$. For these analyses, I only consider the data points where user u posted at least one submission in the subreddit s in H(u, s)$_{\text{future}}$ since my focus was on distinguishing cases where removals occur from cases where there are no removals.

Observation 5: High past removal rate for a user is associated with higher odds of that user experiencing a post removal in the future.

Table 7.6 reports the results of several binomial regression models predicting whether a removal will occur. I began by creating a model B.1 that includes all the subreddit and post history variables as well as the past removal rate (Table 7.4). This model shows that the net subreddit removal rate is associated with higher odds of future removals (OR = 2.461). This suggests the expected association that subreddits that are stricter in their moderation are more likely to remove future postings regardless of the user in consideration. My results also show that a standard deviation increase in the specific past removal rate for each user u in subreddit s leads to a two-fold increase in the odds of future removals (OR = 1.968). Thus, users who have faced more prior removals are likely to have a higher chance of facing a removal again.

Users were more likely to have their posts removed if they submitted in a subreddit that receives more submissions in total (OR = 1.168). One explanation for this is that subreddits that receive many submissions are likely to have a greater number of overall removals. However, posting in a subreddit with a higher number of subscribers was associated with lower odds of future post removals (OR = 0.910).

I found a positive Pearson correlation of statistical significance (r = 0.366, p < .001) between the number of past submissions and future submissions. This high correlation suggests that as the number of past submissions increase, users are also more likely to submit more posts in the future, increasing the likelihood that a future removal will occur if at least one of those future posts is removed. Indeed, I found an odds ratio of 1.164 for past submissions, indicating that a standard deviation increase in the number of past submissions by a user in a subreddit was associated with 16.4% higher odds (100 * (1.164 - 1)) of future removals for the user in that subreddit. Other control variables had much smaller effects on future removals.

Observation 6: Greater explanation rates characterize reduced odds of post removals in the future.

Model B.2 adds the influence of explanation rate to future removals. This model includes only the cases where the user has received at least one post removal in the past and has submitted at least one post in the future. It shows the encouraging result that the odds of the occurrence of future removals lower by 6.5% (OR = 0.935) with each standard deviation increase in the explanation rate. This suggests that explanations help users understand their mistakes and learn the social norms of the community, enabling them to subsequently post submissions that are less likely to get removed.

This result has different implications for different types of communities. For example, consider a small community that experiences 100 post removals a month. Assuming that the odds ratios of model B.2 apply to this community, if explanations rate is increased by two standard deviations, this community may have 2 * 6.5 = 13% fewer post removals or about 87 post removals per month in the future. In contrast, the same increase in explanations rate would cause a large community that usually experiences 10,000 post removals a month to have 1,300 fewer post removals a month in the future. Therefore, moderators on different communities must judge whether the reduction in post removals are worth the investments made in providing removal explanations on their community.

Observation 7: Having a higher fraction of explanations offered through comments, rather than through flairs, is associated with a decreased likelihood of users experiencing a post removal in the future.

Next, I developed a model B.3 to understand the effects of different aspects of explanations on future removals. I found that the average explanation length did not have any significant effect on the occurrence of future removals. One possibility is that as long as explanations provide users the specific information that helps them understand why the removal occurred, the comprehensiveness of explanations do not add to their usefulness. However, I found an odds ratio of 0.905 for explanations through comments rate, indicating that a one unit increase in the rate of explanations provided through comments, rather than through flairs, resulted in a 9.5% decrease (100 * (1 - 0.905)) in the odds of future removals.

Finally, I developed a model B.4 to analyze the impact of explanation source. I found that explanations by bot rate did not have any statistically significant effect on future removals (OR = 1.019). This indicates that the source of removal explanations does not seem to have any substantial effect on the quality of subsequent posts. My interviews with Reddit moderators provides one possible explanation for this. I have found that many moderators use pre-configured removal explanations in order to expedite moderation tasks. Thus, the text outputs for explanations look quite similar, whether they are provided by a human moderator or an automated tool. This may be the reason why the users seem to have similar responses to both human and bot explanations.

While the above analyses evaluate the effects of moderation and explanations on user behaviors across all subreddits, these analyses do not sufficiently take into account the heterogeneity of different Reddit communities because my 'Subreddit variables' only control for a few important factors that distinguish subreddits. In an effort to address this, I used the approach described above to evaluate the effects of explanation rates on future user behaviors in a few individual subreddits. Through these analyses, I also demonstrate how moderators of any community can adopt my approach to evaluate the effects of explanations on that community.

For this, I filtered large, active subreddits (# subscribers > 1M, # submissions > 10K) where explanations are frequently provided (average explanation rate > 0.2). I found four subreddits that satisfied these criteria - *r/politics*, *r/pics*, *r/mildlyinteresting*, and *r/buildapc*. Taking each of these four subreddits one at a time, I built a corpus that only contained <user, subreddit> pairs belonging to that subreddit and developed regression models that test the effects of explanation rates on future postings and future removals on the subreddit. Table 7.7 shows the results of these analyses. Note that I do not include subreddit variables

Table 7.7: Odds ratio of predicting (1) whether the user will post in the future and (2) whether the user's post will get removed in the future on four large, active subreddits. Here, p<0.001: ***; p<0.01:**; p<0.05:*. Each model was constructed on the corpus of the corresponding subreddit for which all the included variables have valid entries. The results show that on each subreddit, higher *explanation rate* are linked to a decrease in the odds of user experiencing a post removal in the future. Other variables are used as controls.

Subreddit	**r/politics**		**r/pics**		**r/mildlyinteresting**		**r/buildapc**	
Dependent Var.	**Future subm.**	**Future removal**	**Future subm.**	**Future removal**	**Future subm.**	**Future removal**	**Future subm.**	**Future removal**
Past Submissions	5022 ***	13.68 ***	2.181 ***	1.217 ***	3.108 ***	1.778 **	1.119	1.048
Avg Past score	0.972	0.889	1.287	1.229	1.12	0.902	0.684	1.133
Avg Past comments	1.04	1.139	0.846	0.834	0.926	1.28	1.56	1.224
Past Removal Rate	0.538 ***	1.521 ***	0.586 ***	1.756 ***	0.771 ***	1.36 ***	0.779 ***	1.331 **
Explanation Rate	0.997	0.877 *	0.943	0.561 ***	1.02	0.795 ***	0.774 *	0.48 ***
Intercept	7.304 ***	7.367 ***	1.317	0.481 ***	0.536 ***	0.431 ***	2.329 *	0.052 ***
Nag. R Square	0.382	0.084	0.204	0.21	0.104	0.049	0.188	0.224
# Obs	4357	2537	4105	1404	4670	1368	419	174

in these analyses as all the data used in each model belong to the same subreddit. These results show that while increases in explanation rates do not significantly affect future submissions on every subreddit, they characterize reduced odds of post removals in the future in every case. This again suggests the important role that explanation mechanisms can play in improving the quality of user contributions.

7.6 Discussion

Online communities thrive on user-generated content. However, inappropriate posts distract from useful content and result in a poor user-experience. Therefore, moderation systems usually desire to increase the number of overall contributions while lowering the number of posts that need to be removed (Grimmelmann, 2015; Kiesler, Kraut, and Resnick, 2012). My analyses in the previous section explored how moderation decisions affect the occurrence of future submissions (Section 7.5.2). I also investigate how moderation actions shape the level of future removals (Section 7.5.3). In this section, I discuss the implications of my results for moderators, site managers, and designers of moderation tools.

7.6.1 Removal Explanations Help Users Learn Social Norms

In prior research, Kiesler et al. have suggested that people learn the norms of a community by (1) Posting and directly receiving feedback, (2) Seeing community guidelines and (3) Observing how other people behave and the consequences of that behavior (Kiesler, Kraut, and Resnick, 2012). I contend that explanations serve as learning resources for Reddit users in each of these three ways.

First, posters who receive explanation messages receive direct feedback from the moderator team in these messages. This can help them realize how their submission did not align with the norms of the community. Therefore, receiving explanations can be a moment of learning for the post submitters.

Second, as my topic modeling and n-gram analyses show, explanations messages usually mention the rule that the submitter has broken (Tables 7.1 and 7.2). For example, topics for explanation comments like "Ask questions in post title" and high frequency of flairs like "non whitelisted domain" and "low effort meme" signify a focus on educating users about the community guidelines. Explanation messages often contain a link to the wiki page for the subreddit rules. Therefore, receiving these explanations increases the likelihood that the moderated users will attend to the community guidelines, and it would help them better understand the explicit social norms of the community[4].

Third, a noteworthy aspect of explanation messages is that they are posted publicly. Although submissions that are removed stop appearing on the front page of the subreddit, they are still accessible to the users who have already engaged with them, for example, through replying via a comment to those submissions. Therefore, many users can still see the removal explanations provided by the moderators. Observing the removal of the post and a reasoned explanation for that removal can inform these bystanders why certain types

[4] Related to this, it is important to consider whether, where, and how prominently community guidelines are posted in discussion spaces. Because certain Reddit interfaces (e.g., mobile website and some third-party Reddit apps) obscure the presence of these guidelines, they may interfere with users' ability to learn the social norms.

of posts are unacceptable on the community. In this way, such interactions can help these bystanders become better content submitters themselves in the future.

7.6.2 Removal Explanations are Linked to Reduction in Post Removals

My regression analyses (Section 7.5) show that when moderated users are provided explanations, their subsequent post removal rate decreases (Observation 6, Model B.2, Table 7.6). As I show in Table 7.7, this relationship holds even in individual subreddits. These are encouraging results, as they indicate that explanations play a role in improving users' posting behaviors. This also raises an interesting question: what would happen to the quality of posted content if 100% of removals were provided explanations? I calculated a rough estimate for this based on my regression results, assuming that the relationship between explanation rate and future removals (Table 7.6) remains linear over a long interval, and noting that an explanation rate of 100% is about 3.20 standard deviations away from mean explanation rate. My calculation shows that the odds of future post removals would reduce by 20.8% if explanations were required to be provided for all removals. Thus, offering explanations could result in a much reduced workload for the moderators.

My LDA analysis of explanation comments (Section 7.3) shows that removal explanations are not just a mechanism to inform users about why the current removal occurred, they are also a means through which moderators can begin to develop a relationship with moderated users. I frequently saw these explanation messages thanking the submitter for posting on the community or expressing regret that the submission had to be removed. This suggests that some of these explanations may have been designed to reduce the moderated users' displeasure about the post removals. Such attempts to engage with the user, in addition to the knowledge about social norms that explanation messages provide, could explain why users who are offered removal explanations submit improved posts in the future.

Prior research has found that in the absence of authoritative explanations, users make sense of content moderation processes by developing "folk theories" about how and why their content was removed (Eslami et al., 2015b). These folk theories often pinpoint to human intervention, including the perceived political biases of moderators, as the primary cause of content removals (West, 2018). My findings suggest that removal explanations can address some of these problems by providing transparency about the moderation mechanisms that shape content removals. For example, the occurrence of LDA topics like "Removal is automatic" suggest an attempt by the moderators to clarify that the post removal was made through the use of automated moderation tools and did not involve human intervention. This increased transparency may contribute to improved user attitudes about the community and motivate users to submit valuable contributions.

I also found that only 1,421 subreddits, a small proportion (0.6%) of all Reddit communities in my data, chose to provide removal reason messages. Thus, explanations are an underutilized moderation mechanism, and site managers should encourage moderators to offer explanations for content removals. Providing explanations may also communicate to the users that the moderator team is committed to providing transparency and being just in their removals.

7.6.3 How should Removal Explanations be Provided?

As I discussed, Reddit moderators can provide explanation messages in a variety of ways. They can comment on a removed submission or flair it. They may compose an explanation message themselves or they may configure a bot to do it. Do these differences matter? Is one approach better than the others in improving future outcomes?

Comments v/s Flairs

My analyses suggest that offering explanation through comments, rather than through flairs, is associated with a decreased likelihood of users experiencing a post removal in the future (Observation 7, Model B.3, Table 7.6). In a similar vein, in Observation 3 (Model A.3, Table 7.5), I note that controlling for explanation rate among other variables, explanation through comments rate is associated with increased odds of future posting.

My findings in Section 7.3 provide clues to interpret these results. Explanation comments differ from flairs in that they cushion against the dissatisfaction resulting from post removals. They often provide information that is future-oriented and go beyond the context of the current removal. Frequently, explanation comments contain information about how users can appeal to reverse the moderation decisions in case the users consider the post removal a mistake. Explanation flairs, on the other hand, usually are very direct, do not employ hedging phrases as frequently, and only pertain to the current removal. These differences may contribute to a relationship between higher levels of explanations through comments and lower levels of future removals.

Although my regression analyses establish the effectiveness of explanation comments over explanation flairs, my data show that flairs are used much more frequently than comments to provide explanations (Section 7.3). This may be because the flairs are much shorter, and therefore, easier for the moderators to provide than comments. Yet, my findings suggest that it may be worthwhile for Reddit moderators to take the time to provide elegant explanations for content removals through comments rather than tagging the post with a short flair. At a broader level, these results indicate that conducting amiable, individualized correspondence with moderated users about their removed posts may be an effective approach for content moderators to nurture potential contributors.

Human moderators v/s automated tools

On the other hand, my results show that controlling for other factors, explanations provided by automated tools or bots are associated with higher odds of moderated users posting in the future (Observation 4, Model A.4, Table 7.5). Additionally, explanations provided by human moderators did not have a significant advantage over explanations provided by bots (Model B.4, Table 7.6) for reducing future post removals. One possible reason for this is that explanations provided by bots are carefully pre-crafted to convey the removal information in a polite manner to the users but spontaneous, impromptu remarks provided by human moderators may not be as polished. This may be why explanations provided by bots result in greater future user activity.

These results suggest an opportunity for deploying automated tools at a higher rate for the purpose of providing explanations. I expect that the field of explainable AI can provide valuable insights for improving the quality of explanations provided by automated tools.

Using these tools can also help address the challenges of scale. When communities grow large quickly and the moderation resources run scarce, it may be difficult for moderators to focus on providing explanations as they are instead engaged in the primary task of firefighting against bad posts. However, if the moderators set up automated tools to provide removal reasons automatically, those tools can continue to provide explanations to users even in high-traffic circumstances.

At the same time, I caution that automated tools should be used with care for the purpose of offering explanations, and in cases where the removal reasons are unclear, human moderators should continue to provide such explanations. As Chapter 5 shows, an overuse of automated tools for content moderation results in cases where these tools make mistakes, thereby causing users to become dissatisfied with the moderation processes. I expect that inaccurate removal explanations are likely to increase resentment among the moderated users rather than improve their attitudes about the community. Therefore, automated tools for providing explanations should be carefully designed and deployed, and their performance should be regularly examined.

7.6.4 When should Removal Explanations be Provided?

My Observation 2 (Model A.2, Table 7.5) states that greater explanation rates are associated with reduced odds of posting in the future. One possible reason for this is that explanations may bring users' attention to the fact that their post has been removed, which they otherwise may not have known about, owing to the frequent silent removals on Reddit. Thus, drawing attention to the removals by providing explanations may irritate users and reduce their user activity. On the other hand, Observation 6 (Model B.2, Table 7.6) notes that providing removal explanations is linked to lower number of future removals. Thus, although offering removal explanations may alienate some users and reduce the likelihood of their future contributions on the community, it may improve the quality of future submissions that *do* get submitted. Therefore, in determining explanation policies, moderators may need to consider whether having high traffic is more important to them than having quality content on their community.

Related to this, it is necessary to consider: In which cases is it worthwhile to provide removal explanations? Should moderators offer an explanation message for every post removal? Or should submissions or post submitters be categorized such that explanations are provided only for certain categories but not others?

It is unclear how providing explanation messages for removing content that is blatantly offensive or trollish would affect the activity of its submitters. As Observation 5 of my Findings (Model B.1, Table 7.6) notes, high past removal rate for a user is associated with higher odds of that user experiencing a post removal in the future, regardless of other factors. It may very well be possible that some bad actors may thrive on the attention they receive for their bad posts from the moderators, and further increase the rate at which they post unacceptable content. Yet, it is difficult to draw the line between inappropriate content that deserves explanations and blatantly offensive content that does not merit providing an explanation. This boundary may also vary between different communities, depending on their social norms, topic and size, among other factors. Furthermore, it may be problematic to classify certain users as irredeemable and unworthy of providing explanation. Thus, significant challenges remain in determining when to use the limited moderated resources in offering explanation mechanisms. I suggest that future research should explore how distinguishing between good actors and bad actors (along a number of dimensions) when providing explanations affects the user-activity and post-removal outcomes.

7.6.5 Limitations and Future Work

As I discussed throughout Section 7.4, I have made many assumptions and simplifications to arrive at the statistical models used in my analyses. I hope that future research in this space starts to inspect these assumptions and explores the role that other factors in moderation systems play in mediating user behaviors. I have only looked at responses to removals that were publicly posted on Reddit communities. It is, however, possible that some subreddits notify users about their content removal through private messages. I only focused on analyzing transparency in regulation of submissions. However, subreddits may also be implementing different levels of transparency in comment removals. It would be useful to focus on moderation of comments in future research.

It is a limitation that this research does not divide users into people we want to post again (well-meaning users who need to be educated in the rules of the community) and people we don't want to post again (users who are being deliberately disruptive, i.e. trolls). Of course, determining who is a troll is subjective and difficult to operationalize fairly (Chapter 3). However, in future work, I would like to separate them if possible to determine what aspects of removal explanations encourage trolls to go away and others to come back. In a similar vein, it would be useful to divide explanations into different categories based on what moderators intended to achieve through those explanations. The topic analyses I presented in Section 7.3 could be a valuable guide to categorize explanations and pursue this direction. I cannot be sure whether users actually read the removal explanations they are given. In future work, I would like to control for this variable.

My large-scale data analysis provides useful insights into how removal and explanation decisions affect future user activity. However, it is critical to investigate the in-situ practical concerns and constraints under which content moderators work. I call for researchers to study how and why moderators currently provide removal explanations and the conditions under which they work. Understanding the perspectives of moderators and building upon current work, researchers can provide design recommendations that are not just valuable for the communities but also feasible for the moderators to implement.

7.7 Conclusion

The sheer volume of content that gets posted on social media platforms makes it necessary for these platforms to rely on moderation mechanisms that are cheap and efficient. However, at this scale and speed, these mechanisms are bound to make many mistakes. Currently, platforms largely make content moderation decisions in an opaque fashion. This secretiveness causes speculations among end-users who suspect that the platforms are biased in some ways (West, 2018). Would it help platforms to instead be transparent about their processes? Would it improve community outcomes if platforms engage with users and explain the reasoning behind their moderation decisions?

In this chapter, I explore the effects of transparency in moderation decisions on user behavior. This research focuses on one important aspect of transparency in content moderation — the explanations about why users' submissions are removed. My findings show that provision of removal explanations is associated with a reduction in future removals, suggesting that taking an educational, rather than a punitive, approach to content moderation can improve community outcomes. My analysis also indicates that using automated tools to provide removal explanations is a promising approach to design for transparency without unduly increasing the work load of moderators.

CHAPTER 8
CONCLUSION

In previous chapters, I have examined some of the key aspects of content moderation. I have investigated the boundaries between controversial speech and online harassment, highlighted tactics identified by my participants as manifestations of online harassment, and offered a theoretical model for perceptions of controversial speech. I have explored the deficiencies of current moderation systems through examining why users have to resort to using third-party moderation tools, and evaluated the efficacy of such tools. My research has also examined the work that goes into enacting content moderation. I have highlighted the concerns of individuals who take on the difficult job of curating content online, focusing on the ways in which automated tools change the nature of their work. I have provided insights into what fairness in content moderation means from the perspectives of end-users who get moderated online. Finally, I have offered empirical evidence for how increasing transparency in moderation decisions can improve the attitudes and future behaviors of moderated users.

Content moderation is not just an ancillary service that platforms provide. It is constitutional of what platforms do (Gillespie, 2018a). While platforms often present themselves as neutral intermediaries, my research shows how the decisions made by their moderation systems mold the dynamics of participation and public discourse. I hope that this work encourages the readers to reflect on what platforms are and what it means for us as a society to rely on them as our information intermediaries.

I began this dissertation asking the question: Where do moderation systems fail, why do they fail, and how can we design new solutions that address those failures? My research has sought to demystify the internal workings of these systems. I have brought to light the different ways in which existing moderation configurations affect a variety of stakeholders, and presented suggestions for how platforms can design to promote prosocial outcomes. While this thesis has explored some key aspects of content moderation, I have only scratched the surface of this complex, multi-layered practice. Because of a rapidly growing research, media and public interest in this topic, I view content moderation emerging as an important interdisciplinary field of study. In this chapter, I conclude with some open questions that emerge from my findings and suggest directions for future research.

8.1 Addressing Online Harassment and Supporting Marginalized Groups Online

This thesis has conceptualized boundaries between controversial speech and online harassment, and presented tactics identified by end-users as manifestations of online harassment. This opens up a number of exciting directions for future work. For example, how can we computationally identify instances of harassment tactics like brigading and sealioning, and promptly moderate them? Can we use theories and principles from the fields of criminology and victimology (e.g., the restorative justice approach) to better serve end-users who have suffered online harassment? What type of design solutions can encourage users to have constructive discussions across disparate social norms of conversation?

An important goal in this space is to support the presence of marginalized groups online. As Chapter 4 shows, transgender users are among the groups most severely affected by online harassment on Twitter. Still, these users cannot simply leave the site because the online transgender community on Twitter serves as a primary source of social support for many of them. This highlights why it is vital to address the problem of online harassment, especially when it is directed against vulnerable individuals. Going forward, it is vital that we identify the specific content moderation needs of marginalized populations such as LGBT and minority groups, for example, through surveys or interviews of individuals from each group. Building upon this, we should also develop tools and systems that serve the special needs of such groups. We can use FATE (Section 2.1.5) and other theories of fairness and justice to evaluate how different approaches to content moderation affect these vulnerable populations. As I found in Chapter 4, harassed users value messages of support during episodes of online abuse, even from strangers. Considering this, designers and researchers should build and evaluate support systems that can help users who have suffered similar abuse connect to one another, and share their experiences. This line of research will offer design and policy insights to help marginalized groups participate in digital public spheres without fear of being abused.

8.2 Evaluating the Effectiveness of Moderation Mechanisms

Although this thesis has examined a few moderation tools and mechanisms like blocklists on Twitter and Automod and removal explanations on Reddit, many other frequently deployed moderation mechanisms remain unexplored. Take the example of ‘deplatforming’ as a moderation strategy on Twitter. Twitter deplatforms, or in other words ban, controversial public figures who promote anti-social ideas, racism, sexism, conspiracy theories, and misinformation. For example, in recent years, Twitter has deplatformed Alex Jones and Milo Yiannopoulos, who are controversial, far-right personalities with huge followings. What happens when key public figures are deplatformed? How do their followers react? Do they find other individuals to rally around and continue spreading toxic ideologies, or do they reform their behavior? Does the number of individuals who continue posting toxic ideas and the volume of such posts reduce? If so, by how much, and how quickly? Examining such questions can offer theoretical and empirical guidelines to assist social media platforms in executing intervention strategies where they deplatform influential public figures.

To take another example, consider the moderation of entire communities. While this thesis has scrutinized the efficacy of user-level interventions like provision of removal explanations, we know comparatively little about community-level interventions. Multi-community platforms like Reddit and Twitch can choose to moderate entire communities. For example, Reddit can ban a community if it fails to comply with the Reddit guidelines.

Alternatively, Reddit administrators can choose to "quarantine" a community. When Reddit quarantines a community, accessing that community displays a warning that requires users to explicitly opt-in to viewing its content. Although the community continues to exist, it does not appear in Reddit search results and its posts are not promoted on other Reddit pages. For example, in recent months, Reddit has quarantined r/TheRedPill, a community focused on misogynistic discussions. What happens after the quarantine? Does the level of activity reduce or is there a Streisand effect[1] and the community attracts more traffic? Does the rate at which newcomers enter the quarantined community change after the quarantine? Where do existing users migrate to? Do they create new toxic subreddits; do they migrate to other, less toxic communities on Reddit; or do they leave Reddit altogether and move to other sites where their activities are less moderated? What does this all mean for the spread of online radicalization? So far, little prior research (e.g., (Chandrasekharan et al., 2017b)) has focused on how community level interventions affect user dynamics. Examining the efficacy of such interventions is critical to developing a holistic understanding of how platforms can use design friction to inhibit the growth of hateful and radical groups online.

These are just two examples of novel moderation strategies that need further scrutiny. Platforms are constantly innovating new approaches to address the challenges they face. Therefore, a systematic analyses of the implications of such approaches is a rich direction for future research.

8.3 Encouraging Voluntary Compliance with Community Norms

This thesis evaluated the effectiveness of providing removal explanations in improving user behaviors. Continuing this line of work, researchers can consider other ways to encourage users' compliance with online community norms through psychological and economic in-

[1] The term "Streisand Effect" emerged out of an incident where an attempt by the actress Barbra Streisand to restrict online views of her Malibu Mansion on a public website had the paradoxical effect of stimulating public interest and leading to many more views than if she had done nothing (Jansen and Martin, 2015).

centives. For example, we can examine the use of graduated sanctions as a moderation strategy. This idea is inspired by Ostrom's groundbreaking studies of institutions that successfully manage common pool resources. In these studies, Ostrom argues that sanctions that are disproportionate to the offense "may produce resentment and unwillingness to confirm to the rules in the future" (Ostrom, 1990). Adopting this design principle in the context of content moderation, I speculate that stronger sanctions will be perceived as more legitimate if they are applied only after lighter sanctions have proven ineffective. However, it is also likely that bad actors may simply ignore lighter measures and create additional burdens on the already limited moderation resources. For example, as Chapter 6 shows, some moderated users identify themselves as trolls and do not consider improving their behaviors despite receiving persuasive messages from moderators. Therefore, I recommend distinguishing bad actors from good–faith users who accidentally break community rules just because they are unaware of the normative practices of that community. Working closely with content moderators and examining large-scale moderation logs, researchers can seek to identify the circumstances in which a mild but certain punishment is more effective in deterring misbehavior than a severe but uncertain punishment. Such efforts can contribute theoretical and practical insights for efficiently channeling moderation resources into nurturing potential contributors.

More broadly, designers should derive from the rich research literature in psychology, economics, criminology and other social sciences to inform their design choices. Although much of this literature has been developed in the context of offline interactions, it provides guidance about individuals motivations and coping mechanisms as well as the conditions under which individuals and communities become successful (Kraut and Resnick, 2012). Researchers should explore how design alternatives backed by theories from this literature play out in online contexts. At the same time, scholars should caution against technological determinism, and consider the complex ways in which design choices and online social systems shape each other.

Appendices

APPENDIX A

GAMERGATE EXAMPLE : THE BALDUR'S GATE CONTROVERSY

Did gamers give negative reviews to the Baldur's Gate expansion because it is a bad game or because it featured a transgender character? This question lay at the center of Baldur's Gate controversy. On March 31 2016, the 18-year old video game Baldur's Gate received a new expansion titled "Siege of Dragonspear" (BeamDog, 2016). The expansion received a barrage of negative user reviews on game shops like GOG and Steam.

Some reviews focused on problems with the game functionality like in-game bugs, dysfunctional multiplayer and mod incompatibility. Other reviews pointed out that the writing of the game was not up to the standard of the original Baldur's Gate games. Many reviews were also politically charged and criticized the inclusion of a transgender character in the expansion. They complained that the developers had crippled the game's creative strengths by shoehorning a token minority character to push their social justice agenda (Monroe, 2016).

Following this, Beamdog, the studio that developed this new expansion of Baldur's Gate, censored discussions about the expansion on its website (Imgur, 2016) and the official Steam forums. This resulted in a backlash from the gamers. They accused Beamdog of trying to cover up the problems in the game by attributing its poor reviews to extremist gamers. Further, many users on KiA were disappointed to find that the game taunted GamerGate by having one of the characters say: "Reeeeaaally, it's all about ethics in heroic adventuring" (Church, 2016).

Many users on KiA felt that the writers should not have hijacked the franchise by forcing their politics into the game. Some users claimed that they appreciate diversity of characters in games, but did not consider the inclusion of LGBT characters a special or revolutionary idea. The only thing that mattered to them is that the characters are well-written.

One user posted on KiA:

> *"I'm trans and what pisses me off is the way the game does this. The trans character, when talked to, starts speechifying about gender before you're allowed to do anything else. Then, when you're finished, the only two options for reply are both positive and polite, which is incredibly immersion-destroying and completely against the philosophy of D&D/Infinity Engine games. In a game where you can murder almost anyone and everyone, and there's a dialogue option for most alignments, apparently not being nice to trans people would be a step too far. Story taking a back seat to politics."*

Users on the other side of the controversy said that the outrage was disproportionate to the perceived offense. They accused the GamerGate supporters of being transphobic, and believed that it was hypocritical of them to force game developers to remove the transgender character, since GamerGate opposes *censorship* in games and media. A user on GamerGhazi, a popular subreddit that hosts anti-GamerGate discussions commented:

> *"I'm very sad that it is even possible for human beings to carry that much hate. And I am even more saddened that a flourishing artistic medium for expression, like videogames, has become the place for those shitstain[s] to carry their battle against minorities. Fuck them."*

Beamdog released a statement that stated that "some of the negative feedback has focused not on *Siege of Dragonspear* but on individual developers at Beamdog, to the point of online threats and harassment" (Campbell, 2016). On the other hand, GamerGate supporters noted that the media eagerly covered the story from one side only, and portrayed them as an angry mob that harassed the progressive game developers.

In summary, there is no single explanation of the controversy about Siege of Dragonspear. Evidence supports the fact that it is indeed a badly designed game (Monroe, 2016).

Many GamerGate supporters claim that the addition of the dialog line about GamerGate is proof that the developers were deliberately trying to provoke GamerGate supporters to get publicity. On the other hand, some vocal opponents of the game are clearly not supporters of transgender rights. This kind of complexity pervades my dataset in Chapter **??**.

APPENDIX B

SURVEY QUESTIONNAIRE

We asked our survey participants the following questions:

1. How much time did you spend in creating this submission?:

   a. < 1 minute

   b. 1-5 minutes

   c. 6-10 minutes

   d. > 10 minutes

2. Which subreddit was the submission posted to?:

   _

3. What is your Reddit username?

   _

4. Before I started this survey, I noticed that this submission was removed:

   a. Yes

   b. No

5. Before I posted this submission, I suspected that it would be removed:

   a. Strongly agree

   b. Agree

   c. Neutral

   d. Disagree

   e. Strongly disagree

6. (If a or b to Q 5): Why did you think it would be removed? Please explain:

   _

7. I think that the removal was fair:

   a. Strongly agree

   b. Agree

   c. Neutral

   d. Disagree

   e. Strongly disagree

8. Please explain how you felt about the removal:

   _

9. Does the subreddit contain rules in its sidebar?

   a. Yes

   b. No

   c. Unsure

10. (If yes to Q 9): I read the rules of the subreddit before posting:

    a. Strongly agree

    b. Agree

    c. Neutral

    d. Disagree

    e. Strongly disagree

11. (If yes to Q 9): The rules on this subreddit are clear:

    a. Strongly agree

    b. Agree

    c. Neutral

    d. Disagree

    e. Strongly disagree

12. Did you notice a comment, flair or private message indicating why your submission was removed?

    a. Yes

    b. No

13. (If yes to Q12) The subreddit provided the reason for why your submission was removed:

    a. Through a comment to the submission

    b. Through a private message

    c. Through a flair to the removed submission

14. (If yes to Q12) Did the removal reason provide you information that you didn't know before?:

    a. Yes

    b. No

15. (If yes to Q12) The removal reason was provided:

    a. By a human

b. By a bot

c. I am unsure

16. This experience changes how I feel about posting on this subreddit in the future:

    a. Strongly agree

    b. Agree

    c. Neutral

    d. Disagree

    e. Strongly disagree

17. (If not c to Q16) Please explain why you feel differently about posting again:

    _

18. How likely are you to post again on this subreddit after this experience?

    a. Very likely

    b. Likely

    c. Neutral

    d. Not likely

    e. Very unlikely

19. Is there anything else you'd like to tell us about your view of this removal?

    _

20. Which country do you live in?

    _

21. What is your age?

a. < 25 years old

b. 25-34 years old

c. 35-44 years old

d. 45-54 years old

e. 55-64 years old

f. 65-74 years old

g. 75 years or older

h. Prefer not to answer

22. What is the highest level of education you have completed?

a. Less than high school

b. High school graduate (includes equivalency)

c. Some college, no degree

d. Associate degree

e. Bachelor’s degree

f. Master’s degree

g. Doctorate degree

h. Prefer not to answer

23. What is your gender?

a. Female

b. Male

c. Another gender

d. Prefer not to answer

24. Would you be willing to participate in a short follow-up interview?

    a. Yes

    b. No

(If yes to Q 24) Please provide your email ID here so that we can contact you for an interview

_ _

REFERENCES

Ackerman, Mark S (2000). “The intellectual challenge of CSCW: the gap between social requirements and technical feasibility”. In: *Human–Computer Interaction* 15.2-3, pp. 179–203.

Allcott, Hunt and Matthew Gentzkow (2017). “Social Media and Fake News in the 2016 Election”. In: *Journal of Economic Perspectives* 31.2, pp. 211–236. URL: `http://pubs.aeaweb.org/doi/10.1257/jep.31.2.211`.

Alstyne, Marshall van and Erik Brynjolfsson (1996). “Electronic Communities: Global Villages or Cyberbalkanization?” In: *ICIS 1996 Proceedings*. URL: `http://aisel.aisnet.org/icis1996/5`.

Armengol, Eva, Albert Palaudaries, and Enric Plaza (2001). “Individual prognosis of diabetes long-term risks: A CBR approach”. In: *Methods of Information in Medicine-Methodik der Information in der Medizin* 40.1, pp. 46–51.

Ashktorab, Zahra and Jessica Vitak (2016). “Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI ’16*. New York, New York, USA: ACM Press, pp. 3895–3905.

Associated Press (2017). *Unreal when it targets you: Faceless trolls attack online*. URL: `http://molawyersmedia.com/2017/04/14/unreal-when-it-targets-you-faceless-trolls-attack-online-2/`.

Auerbach, David (2016). *If Only AI Could Save Us from Ourselves*. URL: `https://www.technologyreview.com/s/603072/if-only-ai-could-save-us-from-ourselves/`.

Automoderator (2018). *Automoderator - reddit.com*. URL: `https://www.reddit.com/wiki/automoderator`.

BeamDog (2016). *Baldur’s Gate: Siege of DragonSpear*. URL: `https://www.siegeofdragonspea com`.

Becker, Howard and Blanche Geer (1957). “Participant observation and interviewing: A comparison”. In: *Human organization* 16.3.

Ben-Kiki, Oren, Clark Evans, and Brian Ingerson (2005). *YAML Ain’t Markup Language*. URL: `http://yaml.org/spec/1.2/spec.html`.

BigQuery (2018). *Google BigQuery*. URL: https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments.

Binns, Reuben et al. (2017). “Like trainer, like bot? Inheritance of bias in algorithmic content moderation”. In: *International Conference on Social Informatics*. Springer, pp. 405–415.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O’Reilly Media, Inc.

Blackwell, Lindsay et al. (2017). “Classification and Its Consequences for Online Harassment: Design Insights from HeartMob.” In: *PACMHCI* 1.CSCW, pp. 24–1.

Blackwell, Lindsay et al. (2018a). “Classification and its Consequences for Online Harassment: Design Insights from HeartMob.” In: *Proceedings of ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW ’18 Online First)*.

Blackwell, Lindsay et al. (2018b). “Understanding Bad Actors Online”. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, W21.

Blackwell, Lindsay et al. (2018c). “When Online Harassment is Perceived as Justified”. In: *Twelfth International AAAI Conference on Web and Social Media*.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

BlockBot, The (2016). *The Block Bot*. URL: http://www.theblockbot.com.

Bozdag, Engin (2013). “Bias in algorithmic filtering and personalization”. In: *Ethics and information technology* 15.3, pp. 209–227.

Braun, Virginia and Victoria Clarke (2006). “Using thematic analysis in psychology”. In: *Qualitative research in psychology* 3.2, pp. 77–101.

Brennan, Jason (2012). *Libertarianism: What Everyone Needs to Know*. Oxford University Press.

Breton, Albert (2007). *The economics of transparency in politics*. Ashgate Publishing, Ltd.

Bruckman, Amy (2002). “Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet”. In: *Ethics and Information Technology* 4.3, pp. 217–231. URL: http://link.springer.com/10.1023/A:1021316409277.

Bruckman, Amy (2006). "Teaching students to study online communities ethically". In: *Journal of Information Ethics*, p. 82.

Bruckman, Amy, Kurt Luther, and Casey Fiesler (2015). "When Should We Use Real Names in Published Accounts of Internet Research?" In: *Digital Research Confidential: The Secrets of Studying Behavior Online*. Ed. by Eszter Hargittai and Christian Sandvig. URL: `https://books.google.com/books?hl=en{\&}lr={\&}id=d1c1CwAAQBAJ{\&}oi=fnd{\&}pg=PA243{\&}dq=bruckman+luther{\&}ots=MGOqhG5zga{\&}sig=RX7JVR6OHkRiy9Pk40dVm2hXEgY{\#}v=onepage{\&}q=bruckmanluther{\&}f=false`.

Bruckman, Amy et al. (1994). "Approaches to managing deviant behavior in virtual communities". In: *CHI Conference Companion*, pp. 183–184.

Bruckman, Amy S et al. (2018). "Managing Deviant Behavior in Online Communities III". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, panel02.

Bruckman (submitter), Amy (2016). *Hi from CS 6470 Design of Online Communities, Georgia Tech. AMA!: KotakuInAction*. URL: `https://redd.it/47wsos`.

Buni, Catherine (2016). *The secret rules of the internet: The murky history of moderation, and how it's shaping the future of free speech*. URL: `https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech`.

Bunt, Andrea, Joanna McGrenere, and Cristina Conati (2007). "Understanding the utility of rationale in a mixed-initiative system for GUI customization". In: *International Conference on User Modeling*. Springer, pp. 147–156.

Campbell, Colin (2016). *Baldur's Gate studio responds to harassment over trans character*. URL: `http://www.polygon.com/2016/4/6/11380556/baldurs-gate-studio-responds-to-harassment-over-trans-character`.

Canavan, Francis (1984). *Freedom of Expression : Purpose as Limit*. Carolina Academic Press, the Claremont Institute for the Study of Statesmanship, and Political Philosophy, p. 181. ISBN: 0890892695.

Caplan, Robyn (2018). *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches*. URL: `https://datasociety.net/output/content-or-context-moderation/`.

Carenini, Giuseppe and Johanna Moore (1998). "Multimedia explanations in IDEA decision support system". In: *Working Notes of the AAAI Spring Symposium on Interactive and Mixed-Initiative Decision Theoretic Systems*, pp. 16–22.

Carlson, Matt (2018). "Facebook in the news: Social media, journalism, and public responsibility following the 2016 trending topics controversy". In: *Digital Journalism* 6.1, pp. 4–20.

Chancellor, Stevie, Zhiyuan Jerry Lin, and Munmun De Choudhury (2016). "This Post Will Just Get Taken Down: Characterizing Removed Pro-Eating Disorder Social Media Content". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 1157–1162.

Chancellor, Stevie et al. (2016). "# thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, pp. 1201–1213.

Chancellor, Stevie et al. (2017). "Multimodal Classification of Moderated Online Pro-Eating Disorder Content". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. New York, New York, USA: ACM Press, pp. 3213–3226. ISBN: 9781450346559. URL: `http://dl.acm.org/citation.cfm?doid=3025453.3025985`.

Chancellor, Stevie et al. (2019). "A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media". In: *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (Atlanta GA*.

Chandrasekharan, Eshwar et al. (2017a). "The bag of communities: Identifying abusive behavior online with preexisting Internet data". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 3175–3187.

Chandrasekharan, Eshwar et al. (Dec. 2017b). "You Can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech". In: *Proc. ACM Hum.-Comput. Interact.* 1.CSCW, 31:1–31:22. URL: `http://doi.acm.org/10.1145/3134666`.

Chandrasekharan, Eshwar et al. (2018). "The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW, p. 32.

Charmaz, Kathy (2006). *Constructing grounded theory: a practical guide through qualitative analysis*. London. ISBN: 9780761973522. arXiv: `arXiv:1011.1669v3`.

Chen, Adrian (2014). *The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed | WIRED*. URL: `https://www.wired.com/2014/10/content-moderation/`.

Cheng, Justin and Michael S. Bernstein (2015). "Flock: Hybrid Crowd-Machine Learning Classifiers". In: *Proceedings of the 18th ACM Conference on Computer Supported Co-*

operative Work &#38; Social Computing. CSCW '15. Vancouver, BC, Canada: ACM, pp. 600–611. ISBN: 978-1-4503-2922-4. URL: http://doi.acm.org/10.1145/2675133.2675214.

Church, Nate (2016). *Developer's Response to 'Baldur's Gate' Controversy Misses the Point.* URL: http://www.breitbart.com/tech/2016/04/04/developers-response-to-baldurs-gate-controversy-misses-the-point/.

Cialdini, Robert B, Carl A Kallgren, and Raymond R Reno (1991). "A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior". In: *Advances in experimental social psychology*. Vol. 24. Elsevier, pp. 201–234.

Citron, Danielle Keats (2009). "Cyber civil rights". In: *BUL Rev.* 89, p. 61.

— (2014). *Hate crimes in cyberspace*. Harvard University Press.

Citron, Danielle Keats and Mary Anne Franks (2014). "Criminalizing Revenge Porn". In: *Wake Forest Law Review* 49. URL: http://heinonline.org/HOL/Page?handle=hein.journals/wflr49{\&}id=357{\&}div=15{\&}collection=journals.

Clément, Maxime and Matthieu J Guitton (2015). "Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia". In: *Computers in Human Behavior* 50, pp. 66–75.

Coleman, Gabriella (2014a). *Hacker, hoaxer, whistleblower, spy: The many faces of Anonymous*. Verso books.

— (2014b). *Hacker, hoaxer, whistleblower, spy: The many faces of Anonymous*. New York: Verso Books.

Collins, Jerri (2018). *The Top 10 Most Popular Sites of 2018*. URL: https://www.lifewire.com/most-popular-sites-3483140.

Cooper, Robyn M and Warren J Blumenfeld (2012). "Responses to Cyberbullying: A Descriptive Analysis of the Frequency of and Impact on LGBT and Allied Youth". In: *Journal of LGBT Youth* 9.2.

Costanza-Chock, Sasha (2018). "Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice". In: URL: https://ssrn.com/abstract=3189696.

Crawford, Kate and Tarleton Gillespie (Mar. 2016). "What is a flag for? Social media reporting tools and the vocabulary of complaint". In: *New Media & Society* 18.3, pp. 410–428. URL: http://journals.sagepub.com/doi/10.1177/1461444814543163.

danah, danah boyd (2008). "Why Youth Heart Social Network Sites: The Role of Networked Publics in Teenage Social Life". In: *MacArthur Foundation Series on Digital Learning – Youth, Identity, and Digital Media*, pp. 119–142. URL: http://ssrn.com/abstract=1518924.

Daniels, Jessie (2009). *Cyber racism: White supremacy online and the new attack on civil rights*. Rowman & Littlefield Publishers.

De Choudhury, Munmun et al. (2016). "Social Media Participation in an Activist Movement for Racial Equality". In: *Tenth International AAAI Conference on Web and Social Media*.

Deimorz (Submitter) (2012). *AutoModerator - a bot for automating straightforward reddit moderation tasks and improving upon the existing spam-filter : TheoryOfReddit*. URL: https://www.reddit.com/r/TheoryOfReddit/comments/onl2u/automoderator_a_bot_for_automating/.

DeNardis, Laura (2012). "HIDDEN LEVERS OF INTERNET CONTROL". In: *Information, Communication & Society* 15.5, pp. 720–738. URL: http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.659199.

DeNardis, Laura and Andrea M Hackl (2015). "Internet governance by social media platforms". In: *Telecommunications Policy* 39.9, pp. 761–770.

DeNardis, Laura and Andrea M. Hackl (2016). "Internet control points as LGBT rights mediation". In: *Information, Communication & Society* 19.6, pp. 753–770. URL: http://www.tandfonline.com/doi/full/10.1080/1369118X.2016.1153123.

DeVito, Michael A, Darren Gergle, and Jeremy Birnholtz (2017). "Algorithms ruin everything:# RIPTwitter, folk theories, and resistance to algorithmic change in social media". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 3163–3174.

DeVito, Michael A et al. (2018). "How people form folk theories of social media feeds and what it means for how we study self-presentation". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, p. 120.

Dewey, Caitlin (2014). *The only guide to Gamergate you will ever need to read - The Washington Post*. URL: https://www.washingtonpost.com/news/theintersect/wp/2014/10/14/the-only-guide-to-gamergate-you-will-ever-need-to-read/?utm{_}term=.acbffb8aceac.

Diakopoulos, Nicholas and Mor Naaman (2011). “Towards Quality Discourse in Online News Comments”. In: *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*. CSCW ’11. Hangzhou, China: ACM, pp. 133–142. ISBN: 978-1-4503-0556-3. URL: http://doi.acm.org/10.1145/1958824.1958844.

Diakopoulos, Nicholas et al. (2017). “Principles for accountable algorithms and a social impact statement for algorithms”. In: *FAT/ML*.

Dibbell, Julian (1994). “A rape in cyberspace or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society”. In: *Ann. Surv. Am. L.* P. 471.

Dillman, Don A et al. (1978). *Mail and telephone surveys: The total design method*. Vol. 19. Wiley New York.

Dimond, Jill P. et al. (2013). “Hollaback! the role of storytelling online in a social movement organization”. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work - CSCW ’13*. New York, New York, USA: ACM Press, p. 477. ISBN: 9781450313315. URL: http://dl.acm.org/citation.cfm?doid=2441776.2441831.

Donath, Judith S (1999). “Identity and deception in the virtual community”. In: *Communities in cyberspace* 1996, pp. 29–59.

Duggan, Maeve (2014). “Online Harassment”. In: *Pew Internet Project*.

— (2017). “Online Harassment”. In: *Pew Internet Project*.

Ellison, Louise and Yaman Akdeniz (1998). “Cyber-stalking: the Regulation of Harassment on the Internet”. In: *Criminal Law Review* 29, pp. 29–48.

Epstein, Dmitry and Gilly Leshed (2016). “The magic sauce: Practices of facilitation in online policy deliberation”. In: *Journal of Public Deliberation* 12.1, p. 4.

Eslami, Motahhare et al. (2015a). “"I Always Assumed That I Wasn’T Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI ’15. Seoul, Republic of Korea: ACM, pp. 153–162. ISBN: 978-1-4503-3145-6. URL: http://doi.acm.org/10.1145/2702123.2702556.

Eslami, Motahhare et al. (2015b). “I always assumed that I wasn’t really that close to [her]: Reasoning about Invisible Algorithms in News Feeds”. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, pp. 153–162.

Eslami, Motahhare et al. (2016). “First i like it, then i hide it: Folk theories of social feeds”. In: *Proceedings of the 2016 cHI conference on human factors in computing systems*. ACM, pp. 2371–2382.

Eslami, Motahhare et al. (2019). “User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, p. 494.

Fieser, James. and Bradley Harris. Dowden (1995). *The internet encyclopedia of philosophy*. Internet Encyclopedia of Philosophy Pub. URL: `http://www.iep.utm.edu/fallacy/{\#}NoTrueScotsman`.

Fiesler, Casey, Jessica L. Feuston, and Amy S. Bruckman (2015). “Understanding Copyright Law in Online Creative Communities”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*. CSCW ’15. Vancouver, BC, Canada: ACM, pp. 116–129. ISBN: 978-1-4503-2922-4. URL: `http://doi.acm.org/10.1145/2675133.2675234`.

Fiesler, Casey et al. (2018). “Reddit Rules! Characterizing an Ecosystem of Governance”. In: *Twelfth International AAAI Conference on Web and Social Media*, pp. 72–81.

Forte, Andrea and Amy Bruckman (Jan. 2008). “Scaling Consensus: Increasing Decentralization in Wikipedia Governance”. In: *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE, pp. 157–157. URL: `http://ieeexplore.ieee.org/document/4438860/`.

French, Megan and Jeff Hancock (2017). “What’s the folk theory? Reasoning about cybersocial systems”. In:

Fudge, James (2013). *The Penny Arcade Controversy That Will Not Die*. URL: `http://gamepolitics.com/2013/09/06/penny-arcade-controversy-will-not-die/`.

Fung, Archon, Mary Graham, and David Weil (2007). *Full disclosure: The perils and promise of transparency*. Cambridge University Press.

GamerGate Wiki (2016). *The Block Bot*. URL: `http://thisisvideogames.com/gamergatewiki/index.php?title=The%7B%5C_%7DBlock%7B%5C_%7DBot`.

GamerGate Wiki (2017). *GGAutoBlocker - GamerGate Wiki*. URL: http://thisisvideogames.com/gamergatewiki/index.php?title=GGAutoBlocker.

Garcia, Sandra E. (2018). *Ex-Content Moderator Sues Facebook, Saying Violent Images Caused Her PTSD*. URL: https://www.nytimes.com/2018/09/25/technology/facebook-moderator-job-ptsd-lawsuit.html.

Geiger, R Stuart (2016). "Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space". In: *Information, Communication & Society* 19.6.

Geiger, R. Stuart and Aaron Halfaker (2013). "When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes?" In: *Proceedings of the 9th International Symposium on Open Collaboration*. WikiSym '13. Hong Kong, China: ACM, 6:1–6:6. ISBN: 978-1-4503-1852-5. URL: http://doi.acm.org/10.1145/2491055.2491061.

Geiger, R. Stuart and David Ribes (2010). "The Work of Sustaining Order in Wikipedia: The Banning of a Vandal". In: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. CSCW '10. Savannah, Georgia, USA: ACM, pp. 117–126. ISBN: 978-1-60558-795-0. URL: http://doi.acm.org/10.1145/1718918.1718941.

Gerrard, Ysabel (2018). "Beyond the hashtag: Circumventing content moderation on social media". In: *New Media & Society* 20.12, pp. 4492–4511. eprint: https://doi.org/10.1177/1461444818776611. URL: https://doi.org/10.1177/1461444818776611.

Gerstenfeld, Phyllis B. (2013). *Hate Crimes : Causes, Controls, and Controversies*, p. 392. ISBN: 1483321851.

Gill, Lex, Dennis Redeker, and Urs Gasser (2015a). "A human rights approach to platform content regulation." In: *Report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. URL: https://freedex.org/a-human-rights-approach-to-platform-content-regulation/.

— (2015b). "Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights". In:

Gillespie, Tarleton (2014). "The relevance of algorithms". In: *Media technologies: Essays on communication, materiality, and society* 167.

— (2015). "Platforms intervene". In: *Social Media+ Society* 1.1, p. 2056305115580479.

Gillespie, Tarleton (2017a). “Governance of and by platforms”. In: *Sage handbook of social media*.

— (2017b). “Governance of and by platforms”. In: *Sage handbook of social media. London: Sage*.

— (2018a). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

— (2018b). *The Logan Paul YouTube controversy and what we should expect from internet platforms*. URL: https://www.vox.com/the-big-idea/2018/1/12/16881046/logan-paul-youtube-controversy-internet-companies.

Glaser, April (2018). *Want a Terrible Job? Facebook and Google May Be Hiring*. URL: https://slate.com/technology/2018/01/facebook-and-google-are-building-an-army-of-content-moderators-for-2018.html.

Glasgow, Brad. *GamerGate*. Unpublished book.

— (2015). *Challenge accepted: interviewing an Internet #hashtag*. URL: http://gamepolitics.com/2015/08/12/challenge-accepted-interviewing-internet-hashtag/.

Gollatz, Kirsten, Felix Beer, and Christian Katzenbach (2018). “The turn to artificial intelligence in governing communication online”. In:

Gonzales, Joseph A., Casey Fiesler, and Amy Bruckman (2015). “Towards an appropriable cscw tool ecology: Lessons from the greatest international scavenger hunt the world has ever seen”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW ’15*. New York, New York, USA: ACM Press, pp. 946–957.

Gorwa, Robert (2019). “What is platform governance?” In: *Information, Communication & Society*, pp. 1–18.

Granados, Nelson and Alok Gupta (2013). “Transparency strategy: Competing with information in a digital world.” In: *MIS quarterly* 37.2.

Gregor, Shirley and Izak Benbasat (1999). “Explanations from intelligent systems: Theoretical foundations and implications for practice”. In: *MIS quarterly*, pp. 497–530.

Grenoble, Ryan (2013). *Facebook Reverses Stance On Beheading Videos, But Nipples Are Still A No-No (UPDATE) | HuffPost*. URL: https://www.huffingtonpost.

com/2013/10/22/facebook-allows-beheading-videos-graphic-content{_}n{_}4143244.html.

Grevet, Catherine (2016). “Being nice on the internet: designing for the coexistence of diverse opinions online”. PhD thesis. Georgia Institute of Technology.

Grimmelmann, James (2015). “The virtues of moderation”. In: *Yale JL & Tech.* 17, p. 42.

Gross, Ralph and Alessandro Acquisti (2005). “Information revelation and privacy in online social networks”. In: *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pp. 71–80. URL: http://portal.acm.org/citation.cfm?doid=1102199.1102214%20http://portal.acm.org/citation.cfm?id=1102214.

Helberger, Natali, Jo Pierson, and Thomas Poell (2018). “Governing online platforms: From contested to cooperative responsibility”. In: *The information society* 34.1, pp. 1–14.

Herlocker, Jonathan L, Joseph A Konstan, and John Riedl (2000). “Explaining collaborative filtering recommendations”. In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, pp. 241–250.

Herring, Susan et al. (2002). “Searching for safety online: Managing "trolling" in a feminist forum”. In: *The Information Society* 18.5, pp. 371–384. URL: http://www.tandfonline.com/doi/abs/10.1080/01972240290108186.

Herring, Susan C (2000). “Gender differences in CMC: Findings and implications”. In: *Computer Professionals for Social Responsibility Journal* 18.1, p. 0.

Hess, Amanda (2014). *Twitter harassment: User-created apps Block Together, Flaminga, and the Block Bot crack down on Twitter abuse.* URL: http://www.slate.com/blogs/future%7B%5C_%7Dtense/2014/08/06/twitter%7B%5C_%7Dharassment%7B%5C_%7Duser%7B%5C_%7Dcreated%7B%5C_%7Dapps%7B%5C_%7Dblock%7B%5C_%7Dtogether%7B%5C_%7Dflaminga%7B%5C_%7Dand%7B%5C_%7Dthe%7B%5C_%7Dblock.html.

Hochschild, Arlie Russell (1983). *The managed heart: Commercialization of human feeling*. Berkeley: University of California Press.

Hoffmann, Anna Lauren (2019). “Where fairness fails: On data, algorithms, and the limits of antidiscrimination discourse”. In: *Information, Communication, and Society* 22.7.

Holbrook, Allyson L and Jon A Krosnick (2009). “Social desirability bias in voter turnout reports: Tests using the item count technique”. In: *Public Opinion Quarterly* 74.1, pp. 37–67.

Hughey, Matthew W and Jessie Daniels (2013). “Racist comments at online news sites: a methodological dilemma for discourse analysis”. In: *Media, Culture & Society* 35.3, pp. 332–347.

Imgur (2016). *Dee, Baldur’s Gate developer, locks down the Steam forums and continues banning users for criticizing the Baldur’s Gate Expansion*. URL: http://imgur.com/DoNRiVg.

ITSigno (submitter) (2017). *Posting Guidelines replacing Rule 3*. URL: https://www.reddit.com/r/KotakuInAction/comments/5si6cp/posting{_}guidelines{_}replacing{_}rule{_}3/.

Jansen, Sue Curry and Brian Martin (2015). “The Streisand effect and censorship backfire”. In:

Jason, Zachary (2015). *Game of Fear*. URL: http://www.bostonmagazine.com/news/article/2015/04/28/gamergate/.

Jhaver, Shagun, Amy Bruckman, and Eric Gilbert (Nov. 2019). “Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW. URL: https://doi.org/10.1145/3359252.

Jhaver, Shagun, Larry Chan, and Amy Bruckman (2018). “The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action”. In: *First Monday* 23.2. URL: http://firstmonday.org/ojs/index.php/fm/article/view/8232.

Jhaver, Shagun, Yoni Karpfen, and Judd Antin (2018). “Algorithmic Anxiety and Coping Strategies of Airbnb Hosts”. In: *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems*.

Jhaver, Shagun, Pranil Vora, and Amy Bruckman (2017). *Designing for Civil Conversations: Lessons Learned from ChangeMyView*. Tech. rep. Georgia Institute of Technology.

Jhaver, Shagun et al. (Mar. 2018). “Online Harassment and Content Moderation: The Case of Blocklists”. In: *ACM Trans. Comput.-Hum. Interact.* 25.2. URL: https://doi.org/10.1145/3185593.

Jhaver, Shagun et al. (July 2019a). “Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator”. In: *ACM Trans. Comput.-Hum. Interact.* 26.5. URL: https://doi.org/10.1145/3338243.

Jhaver, Shagun et al. (Nov. 2019b). "“Did You Suspect the Post Would Be Removed?”: Understanding User Reactions to Content Removals on Reddit". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW. URL: `https://doi.org/10.1145/3359294`.

Jiang, Ling and Eui-Hong Han (2019). "ModBot: Automatic Comments Moderation". In: *Computation+ Journalism Symposium*.

Kain, Erik (2014). *GamerGate: A Closer Look At The Controversy Sweeping Video Games.* URL: `http://www.forbes.com/sites/erikkain/2014/09/04/gamergate-a-closer-look-at-the-controversy-sweeping-video-games/{\#}52d88b5d5448`.

Kaptelinin, Victor (1996). "Activity theory: Implications for human-computer interaction". In: *Context and consciousness: Activity theory and human-computer interaction* 1, pp. 103–116.

Kelion, Leo (2013). *Facebook lets beheading clips return to social network - BBC News.* URL: `http://www.bbc.com/news/technology-24608499`.

Kerr, Aphra and John D Kelleher (2015). "The recruitment of passion and community in the service of capital: Community managers in the digital games industry". In: *Critical Studies in Media Communication* 32.3, pp. 177–192.

Kiene, Charles, Andrés Monroy-Hernández, and Benjamin Mako Hill (2016). "Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: ACM, pp. 1152–1156. ISBN: 978-1-4503-3362-7. URL: `http://doi.acm.org/10.1145/2858036.2858356`.

Kiesler, Sara, Robert Kraut, and Paul Resnick (2012). "Regulating behavior in online communities". In: *Building Successful Online Communities: Evidence-Based Social Design*.

Kim, Nancy S. (2009). *Website Proprietorship and Online Harassment*. URL: `https://papers.ssrn.com/sol3/papers.cfm?abstract{\_}id=1354466`.

Kizilcec, René F (2016). "How much information?: Effects of transparency on trust in an algorithmic interface". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 2390–2395.

Klein, David A and Edward H Shortliffe (1994). "A framework for explaining decision-theoretic advice". In: *Artificial Intelligence* 67.2, pp. 201–243.

Klonick, Kate (2017). “The New Governors: The People, Rules, and Processes Governing Online Speech”. In: *Harvard Law Review* 131. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=2937985.

Know Your Meme (2016). *Social Justice Warrior*. URL: http://knowyourmeme.com/memes/social-justice-warrior.

Kotaku In Action (2016). *Kotaku in Action: The almost-official GamerGate subreddit*. URL: https://www.reddit.com/r/kotakuinaction/.

KotakuInAction (2015). *How to tell legitimate criticism of games and gaming trends from "anti-game" criticism?* URL: https://redd.it/3081x8.

Kraut, Robert E and Paul Resnick (2012). *Building successful online communities: Evidence-based social design*. MIT Press. URL: https://books.google.it/books?hl=en{\&}lr={\&}id=lIvBMYVxWJYC{\&}oi=fnd{\&}pg=PR7{\&}dq=Building+Successful+Online+Communities:+Evidence-Based+Social+Design{\&}ots=zYFYfmofIA{\&}sig=npdpDyk3PXLgLe8PmffPFLm9T9M.

Kriplean, Travis et al. (2012). “Supporting reflective public thought with considerit”. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW ’12*. Seattle, Washington, USA: ACM Press, pp. 265–274. URL: http://dl.acm.org/citation.cfm?doid=2145204.2145249.

Lakkaraju, Himabindu, Stephen H. Bach, and Jure Leskovec (2016). “Interpretable Decision Sets: A Joint Framework for Description and Prediction”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*. New York, New York, USA: ACM Press, pp. 1675–1684. ISBN: 9781450342322. URL: http://dl.acm.org/citation.cfm?doid=2939672.2939874.

Lampe, Cliff, Erik Johnston, and Paul Resnick (2007). “Follow the Reader: Filtering Comments on Slashdot”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’07. San Jose, California, USA: ACM, pp. 1253–1262. ISBN: 978-1-59593-593-9. URL: http://doi.acm.org/10.1145/1240624.1240815.

Lampe, Cliff and Paul Resnick (2004). “Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*.

Lampe, Cliff et al. (2014). “Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums”. In: *Government Information Quarterly* 31.2, pp. 317 –326. URL: http://www.sciencedirect.com/science/article/pii/S0740624X14000021.

Lenhart, Amanda et al. (2016). *Online Harassment, Digital Abuse, and Cyberstalking in America*. URL: https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/.

Long, Kiel et al. (2017). “"Could You Define That in Bot Terms"?: Requesting, Creating and Using Bots on Reddit”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. Denver, Colorado, USA: ACM, pp. 3488–3500. ISBN: 978-1-4503-4655-9. URL: http://doi.acm.org/10.1145/3025453.3025830.

Lotan, Gilad et al. (2011). “The Arab Spring - the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions.” In: *International journal of communication* 5.

Lwin, May O, Benjamin Li, and Rebecca P Ang (2012). “Stop bugging me: An examination of adolescents’ protection behavior against online harassment”. In: *Journal of Adolescence*. URL: https://www.researchgate.net/profile/May{_}Lwin/publication/51500980{_}Stop{_}Bugging{_}Me{_}An{_}Examination{_}of{_}Adolescents'{_}Protection{_}Behavior{_}Against{_}Online{_}Harrassment/links/5453b48b0cf26d50
pdf.

Madrigal, Alexis (2018). *Inside Facebook’s Fast-Growing Content-Moderation Effort*. URL: https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/.

Malki, David (2014). *The Terrible Sea Lion*. URL: http://wondermark.com/1k62/.

Mark, Gloria, Yiran Wang, and Melissa Niiya (2014). “Stress and Multitasking in Everyday College Life: An Empirical Study of Online Activity”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’14. Toronto, Ontario, Canada: ACM, pp. 41–50. ISBN: 978-1-4503-2473-1. URL: http://doi.acm.org/10.1145/2556288.2557361.

Martin, Fiona (2015). *Getting my two cents worth in: Access, interaction, participation and social inclusion in online news commenting*.

Matias, J Nathan (2016a). “Going dark: Social factors in collective action against platform operators in the Reddit blackout”. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, pp. 1138–1151.

Matias, J Nathan et al. (2015). “Reporting, reviewing, and responding to harassment on Twitter”. In:

Matias, Nathan J. (2016b). *Posting Rules in Online Discussions Prevents Problems & Increases Participation.* URL: https://civilservant.io/moderation{_}experiment{_}r{_}science{_}rule{_}posting.html.

— (2016c). "The Civic Labor of Online Moderators". In: *Internet Politics and Policy conference*. Oxford, United Kingdom.

— (2018). *Gathering the Custodians of the Internet: Lessons from the First CivilServant Summit.* URL: https://civilservant.io/civilservant_summit_report_jan_2018.html.

— (2019). "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions". In: *Proceedings of the National Academy of Sciences* 116.20, pp. 9785–9789.

McDaniel, Patrick, Nicolas Papernot, and Z. Berkay Celik (2016). "Machine Learning in Adversarial Settings". In: *IEEE Security & Privacy* 14.3, pp. 68–72. URL: http://ieeexplore.ieee.org/document/7478523/.

McGillicuddy, Aiden, Jean-Gregoire Bernard, and Jocelyn Cranefield (2016). "Controlling Bad Behavior in Online Communities: An Examination of Moderation Work". In: *ICIS 2016 Proceedings*. URL: http://aisel.aisnet.org/icis2016/SocialMedia/Presentations/23.

Meditations, Untimely (2016). *A People's History of GamerGate*. URL: http://www.historyofgamergate.com.

Melendez, Steven (2015). *Here's How 20,000 Reddit Volunteers Fight Trolls, Spammers, And Played-Out Memes*. URL: https://www.fastcompany.com/3048406/heres-how-20000-reddit-volunteers-fight-trolls-spammers-and-played-out-memes.

Merriam, Sharan B (2002). "Introduction to Qualitative Research". In: *Qualitative research in practice: Examples for discussion and analysis* 1.

Monroe, Nick (2016). *Trying to Understand Baldur's GateGate: The Controversy Explained.* URL: http://gameranx.com/features/id/47159/article/trying-to-understand-baldurs-gategate/.

Moreau, Elise (2016). *The Top 25 Social Networking Sites People Are Using*. URL: https://www.lifewire.com/top-social-networking-sites-people-are-using-3486554.

— (2017). *What Exactly Is a Reddit AMA?* URL: https://www.lifewire.com/what-exactly-is-a-reddit-ama-3485985.

Morris, Kevin (2015). *Reddit moderation being taken over by bots-and that's a good thing*. URL: https://www.dailydot.com/news/reddit-automoderator-bots/.

Mortensen, Torill Elvira (2016). "Anger, Fear, and Games The Long Event of #GamerGate". In: *Games and Culture*, p. 1555412016640408.

Moser, Cornelia (2001). "How open is' open as possible'?: three different approaches to transparency and openness in regulating access to EU documents". In:

Müller, Hendrik, Aaron Sedley, and Elizabeth Ferrall-Nunge (2014). "Survey research in HCI". In: *Ways of Knowing in HCI*. Springer, pp. 229–266.

Mumford, Enid (2000). "A socio-technical approach to systems design". In: *Requirements Engineering* 5.2, pp. 125–133.

Nagle, Angela (2017). *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing.

Newman, David et al. (2010). "Automatic evaluation of topic coherence". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 100–108.

Newman, Nic (2009). "The rise of social media and its impact on mainstream journalism". In:

Nieborg, David B and Thomas Poell (2018). "The platformization of cultural production: Theorizing the contingent cultural commodity". In: *new media & society* 20.11, pp. 4275–4292.

Nunziato, Dawn C (2005). "The Death of the Public Forum in Cyberspace". In: *Berkeley Technology Law Journal* 20.2. URL: http://dx.doi.org/doi:10.15779/Z38039W.

Oetheimer, Mario (2009). "Protecting Freedom of Expression: The Challenge of Hate Speech in the European Court of Human Rights Case Law". In: *Cardozo Journal of International and Comparative Law* 17. URL: http://heinonline.org/HOL/Page?handle=hein.journals/cjic17{\&}id=433{\&}div={\&}collection=.

Ong, Erica (2018). *Is Machine Learning the Future of Content Moderation?* URL: https://insights.conduent.com/conduent-blog/is-machine-learning-the-future-of-content-moderation.

Opazo, M Pilar (2010). "Revitalizing the Concept of Sociotechnical Systems in Social Studies of Technology". In:

Ostrom, Elinor (1990). *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press, Cambridge.

Park, Deokgun et al. (2016). "Supporting comment moderators in identifying high quality online news comments". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 1114–1125.

Pater, Jessica A. et al. (2016). "Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms". In: *Proceedings of the 19th International Conference on Supporting Group Work*. GROUP '16. Sanibel Island, Florida, USA: ACM, pp. 369–374. ISBN: 978-1-4503-4276-6. URL: http://doi.acm.org/10.1145/2957276.2957297.

Patreon (2017). *Randi Harper is creating Online Activism and Open Source Anti-Harassment Tools | Patreon*. URL: https://www.patreon.com/freebsdgirl.

Patton, Michael Quinn (1990). *Qualitative evaluation and research methods*. SAGE Publications, Inc.

Peña Gangadharan, Seeta and Jędrzej Niklas (2019). "Decentering technology in discourse on discrimination". In: *Information, Communication & Society* 22.7, pp. 882–899.

Phillips, Whitney (2015a). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press.

— (2015b). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Pierson, Emma (2015). "Outnumbered but well-spoken: Female commenters in the New York Times". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, pp. 1201–1213.

Pine, Dan (2016). *Anti-Semitic emails, tweets hit candidates and journalists*. URL: http://www.jweekly.com/2016/06/17/anti-semitic-emails-tweets-hit-candidates-and-journalists/.

Postman, Neil (1992). *Technopoly*. New York: Vintage.

PRAW (2016). *The Python Reddit API Wrapper*. URL: http://praw.readthedocs.io/en/stable/.

Pu, Pearl and Li Chen (2006). “Trust building with explanation interfaces”. In: *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, pp. 93–100.

Rader, Emilee, Kelley Cotter, and Janghee Cho (2018). “Explanations as mechanisms for supporting algorithmic transparency”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, p. 103.

RationalWiki (2016). *Atheism Plus*. URL: http://rationalwiki.org/wiki/Atheism%7B%5C_%7DPlus.

Rawlins, Brad (2008). “Give the emperor a mirror: Toward developing a stakeholder measurement of organizational transparency”. In: *Journal of Public Relations Research* 21.1, pp. 71–99.

RedditBots (2019). *autowikibot*. URL: https://www.reddit.com/r/autowikibot/wiki/redditbots.

Reddit.com (2016). *Content policy*. URL: https://www.reddit.com/help/contentpolicy.

Renfro, Kim (2016). *For whom the troll trolls: A day in the life of a Reddit moderator*. URL: https://www.businessinsider.com/what-is-a-reddit-moderator-2016-1#crocker-has-.

Roberts, Sarah (2016). “Commercial Content Moderation: Digital Laborers’ Dirty Work”. In: *Media Studies Publications*. URL: https://ir.lib.uwo.ca/commpub/12.

Roberts, Sarah T. (2014). “Behind the screen: the hidden digital labor of commercial content moderation”. PhD thesis. University of Illinois at Urbana-Champaign. URL: https://www.ideals.illinois.edu/handle/2142/50401.

Rode, Jennifer A. (2011). “Reflexivity in digital anthropology”. In: *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI ’11*. New York, New York, USA: ACM Press, p. 123. ISBN: 9781450302289. URL: http://dl.acm.org/citation.cfm?doid=1978942.1978961.

Romano, Aja (2017). *How the alt-right uses internet trolling to confuse you into dismissing its ideology*. URL: https://www.vox.com/2016/11/23/13659634/alt-right-trolling.

Ronson, Jon (2015). *How One Stupid Tweet Blew Up Justine Sacco’s Life*. URL: http://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html?{_}r=0.

Ruiz, Rebecca (2014). *When Your Job Is to Moderate the Internet's Nastiest Trolls*. URL: http://mashable.com/2014/09/28/moderating-the-trolls/%7B%5C#%7Dj.2EPKo8mkqc.

Saha, Koustuv, Ingmar Weber, and Munmun De Choudhury (2018). "A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses". In: *ICWSM*.

Saha, Koustuv et al. (2019). "A Social Media Study on The Effects of Psychiatric Medication Use". In: *ICWSM*.

Schesser, Stacey D. (2006). "A New Domain for Public Speech: Opening Public Spaces Online". In: *California Law Review* 94.6, p. 1791. URL: http://www.jstor.org/stable/10.2307/20439081?origin=crossref.

Schneider, Jodi, Tudor Groza, and Alexandre Passant (2013). "A review of argumentation for the social semantic web". In: *Semantic Web* 4.2, pp. 159–218.

Schrock, Andrew and danah Boyd (2011). "Problematic youth interaction online: Solicitation, harassment, and cyberbullying". In: *Computer-mediated communication in personal relationships*.

Scollon, Christie Napa, Chu-Kim Prieto, and Ed Diener (2009). "Experience sampling: promises and pitfalls, strength and weaknesses". In: *Assessing well-being*. Springer, pp. 157–180.

Scott, Mark and Mike Isaac (2016). "Facebook restores iconic Vietnam War photo it censored for nudity". In: *The New York Times*.

Seering, Joseph, Robert Kraut, and Laura Dabbish (2017). "Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA: ACM, pp. 111–125. ISBN: 978-1-4503-4335-0. URL: http://doi.acm.org/10.1145/2998181.2998277.

Seering, Joseph et al. (2019a). "Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, p. 606.

Seering, Joseph et al. (2019b). "Moderator engagement and community development in the age of algorithms". In: *New Media & Society*, p. 1461444818821316.

Singh, Monika, Divya Bansal, and Sanjeev Sofat (2016). "Behavioral analysis and classification of spammers distributing pornographic content in social media". In: *Social Network Analysis and Mining* 6.1, p. 41.

Sjwomble (2016). *The Problem With Personal Block Lists*. URL: https://sjwomble.wordpress.com/2016/04/28/the-problem-with-personal-block-lists/.

Smith, Aaron and M Anderson (2018). "Social media use in 2018. Pew Research Center [Internet]". In: *Science & Tech. URl: http://www. pewinternet. org/2018/03/01/social-media-usein-2018/(visited on 04/16/2018).*

Soni, Devin and Vivek K. Singh (Nov. 2018). "See No Evil, Hear No Evil: Audio-Visual-Textual Cyberbullying Detection". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW, 164:1–164:26. URL: http://doi.acm.org/10.1145/3274433.

Spender, Dale (1985). *Man made language*. Pandora.

Squirrell, Tim (2019). "Platform dialectics: The relationships between volunteer moderators and end users on reddit". In: *New Media & Society*, p. 1461444819834317.

Statista (2017). *Twitter: number of active users 2010-2017*. URL: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

Stecklow, Steve (2018). *Why Facebook is losing the war on hate speech in Myanmar*. URL: https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

Straus, Anselm and Juliet Corbin (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*.

(Submitter), captainmeta4 (2016). *What is /u/BotBust? : BotBust*. URL: https://www.reddit.com/r/BotBust/comments/5092dg/what{_}is{_}ubotbust/.

Suzor, Nicolas (2018). "Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms". In: *Social Media + Society* 4.3, p. 2056305118787812.

Suzor, Nicolas, Tess Van Geelen, and Sarah Myers West (2018). "Evaluating the legitimacy of platform governance: A review of research and a shared research agenda". In: *International Communication Gazette* 80.4, pp. 385–400.

Suzor, Nicolas P et al. (2019). "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation". In: *International Journal of Communication* 13, p. 18.

Swearingen, Courtnie and Brian Lynch (2018). *We're Reddit Mods, and This Is How We Handle Hate Speech*. URL: https://www.wired.com/2015/08/reddit-mods-handle-hate-speech/.

Taylor, Linnet (2017). “What is data justice? The case for connecting digital rights and freedoms globally”. In: *Big Data & Society* 4.2, p. 2053951717736335.

Taylor, Steven J and Robert Bogdan (1998). *Introduction to qualitative research methods: The search for meaning*. Third. John Wiley & Sons.

Taylor, Steven J, Robert Bogdan, and Marjorie DeVault (2015). “Participant Observation: In the Field”. In: *Introduction to qualitative research methods: A guidebook and resource*. John Wiley & Sons. Chap. 3.

Taylor, TL (2018). “Regulating the networked broadcasting frontier”. In: *Watch me play: Twitch and the rise of game live streaming*. Princeton University Press. Chap. 5.

Thompson, Ken (1968). “Programming Techniques: Regular expression search algorithm”. In: *Communications of the ACM* 11.6, pp. 419–422. URL: http://portal.acm.org/citation.cfm?doid=363347.363387.

Tiku, Nitasha and Casey Newton (2015). *Twitter CEO: ‘We suck at dealing with abuse’ - The Verge*. URL: http://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the.

Titley, Gavan, Ellie Keen, and László Földi (2014). “Starting points for combating hate speech online”. In: *Council of Europe, October 2014*.

Treré, Emiliano (2012). “Social movements as information ecologies: Exploring the coevolution of multiple Internet technologies for activism”. In: *International Journal of Communication* 6.

Turkle, S (2006). “Life on the screen: Identity in the age of the internet”. In: URL: http://www.citeulike.org/group/48/article/949801.

Twitter (2016a). *Blocking accounts on Twitter*. URL: https://support.twitter.com/articles/117063.

— (2016b). *What are replies and mentions?* URL: https://support.twitter.com/articles/14023.

Uscinski, Joseph E, Darin DeWitt, and Matthew D Atkinson (2018). “A Web of Conspiracy? Internet and Conspiracy Theory”. In: *Handbook of Conspiracy Theory and Contemporary Religion*. BRILL, pp. 106–130.

Vanian, Jonathan (2017). *Twitter Toughens Rules on Nudity and Revenge Porn | Fortune*. URL: http://fortune.com/2017/10/27/nudity-revenge-porn-twitter/.

Velden, Theresa and Carl Lagoze (2013). “The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication”. In: *Journal of the American Society for Information Science and Technology* 64.12, pp. 2405–2427.

Vitak, Jessica et al. (2017). “Identifying Women’s Experiences With and Strategies for Mitigating Negative Effects of Online Harassment”. In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW ’17*.

Wagner, Kyle (2012). *The Worst Job at Google: A Year of Watching Beastiality, Child Pornography, and Other Terrible Internet Things*. URL: http://gizmodo.com/5936572/the-worst-job-at-google-a-year-of-watching-beastiality-child-pornography-and-other-terrible-internet-things.

Wallach, Hanna M et al. (2009). “Evaluation methods for topic models”. In: *Proceedings of the 26th annual international conference on machine learning*. ACM, pp. 1105–1112.

Wang, Weiquan and Izak Benbasat (2007). “Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs”. In: *Journal of Management Information Systems* 23.4, pp. 217–246.

Warzel, Charlie (2016). *“A Honeypot For Assholes”: Inside Twitter’s 10-Year Failure To Stop Harassment*. URL: https://www.buzzfeed.com/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s.

West, Sarah Myers (2017). “Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms”. In: *Media and Communication* 5.3, p. 28. URL: https://www.cogitatiopress.com/mediaandcommunication/article/view/989.

— (2018). “Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms”. In: *New Media & Society*.

West (Speaker), Lindy (2015). *If You Don’t Have Anything Nice to Say, SAY IT IN ALL CAPS*. URL: http://www.thisamericanlife.org/radio-archives/episode/545/transcript.

Wiki, GamerGate (2016). *Main Page*. URL: http://thisisvideogames.com/gamergatewiki/index.php.

Wilburn, Vanessa L (1994). “Gender and anonymity in computer-mediated communication: participation, flaming, deindividuation”. PhD thesis. University of Florida.

Wortham, Jenna (2017). *Why Can't Silicon Valley Fix Online Harassment? - The New York Times*. URL: `https://www.nytimes.com/2017/04/04/magazine/why-cant-silicon-valley-fix-online-harassment.html?{\_}r=0`.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon (2017). "Ex Machina: Personal Attacks Seen at Scale". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, pp. 1391–1399. ISBN: 978-1-4503-4913-0. URL: `https://doi.org/10.1145/3038912.3052591`.

Ybarra, Michele L and Kimberly J Mitchell (2004). "Youth engaging in online harassment: associations with caregiver–child relationships, Internet use, and personal characteristics". In: *Journal of Adolescence* 27.3, pp. 319–336.

Young, Cathy (2015). *When Trolls Attack & GamerGate Is Scapegoated*. URL: `http://www.realclearpolitics.com/articles/2015/11/25/when{\_}trolls{\_}attack{\_}{\_}gamergate{\_}is{\_}scapegoated{\_}128844.html`.

Zephoria (2018). *Top 20 Facebook Statistics - Updated March 2018*. URL: `https://zephoria.com/top-15-valuable-facebook-statistics/`.

Zoller, Elisabeth (2009). "Freedom of Expression: "Precious Right" in Europe, "Sacred Right" in the United States? Symposium: An Ocean Apart? Freedom of Expression in Europe and the United States Foreword: Freedom of Expression: "Precious Right" in Eur". In: *Law Journal* 84.2. URL: `http://www.repository.law.indiana.edu/ilj`.

Zúñiga, Homero Gil de, Nakwon Jung, and Sebastián Valenzuela (2012). "Social media use for news and individuals' social capital, civic engagement and political participation". In: *Journal of computer-mediated communication* 17.3, pp. 319–336.