TOWARD A MULTI-STAKEHOLDER PERSPECTIVE FOR IMPROVING ONLINE CONTENT MODERATION

by

JIALUN "AARON" JIANG | 姜嘉伦

B.S., University of Minnesota, 2016
M.S., University of Colorado, 2019

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement of the degree of
Doctor of Philosophy
Department of Information Science
2020

Committee Members:
Dr. Casey Fiesler, University of Colorado
Dr. Jed R. Brubaker, University of Colorado
Dr. Brian Keegan, University of Colorado
Dr. Amy Bruckman, Georgia Institute of Technology
Dr. Eric Gilbert, University of Michigan

abstract

Jiang, Jialun "Aaron" (Ph.D., Information Science)

Toward a Multi-stakeholder Perspective for Improving Online Content Moderation

Thesis directed by Assistant Professor Casey Fiesler and Assistant Professor Jed R. Brubaker

Online communities have struggled with malicious behavior, and a major way to combat such abuse is content moderation. While content moderation has been effective in addressing problems in online communities, it also faces various challenges, on the levels of both larger platforms and smaller communities, and these challenges often arise from the varying needs of different stakeholders. For example, global platforms need to consider different values and cultures while dealing with the massive amount of content waiting to be reviewed, and communities with different technological infrastructures also have needs for different rules, moderation strategies, and tools. My dissertation provides a multi-stakeholder perspective of the challenges of online content moderation, provides actionable guidelines to address problems for both volunteer and commercial moderation, and argues that different stakeholders and their associated trade-offs should be a central consideration in online content moderation.

In my dissertation, I first describe challenges brought by moderating different technologies by examining the challenges that new platform technology brings to community moderation through a case study of moderating real-time voice on Discord, and argue that community moderators and technology designers should cater to the unique technological infrastructures of

individual platforms and communities. My work then investigates the multi-stakeholder tensions in commercial moderation, and reveals varied perceptions of abusive behavior from global social media users, demonstrating the limitations of using a single set of rules to govern global users. Building on my empirical work about the pervasive multi-stakeholder challenges, using a systematic literature review, I propose a framework that centers trade-offs in online content moderation, and show how trade-offs are core to the very definition of content moderation.

This dissertation provides deep, empirical understandings of how multiple stakeholders are involved in content moderation, and how ignoring stakeholders' needs can lead to serious problems and consequences. By contributing a new way to conceptualize content moderation, my work argues for a future where we start to see different stakeholders, acknowledge their needs, and consciously address the trade-offs between their needs.

dedication

For my family and friends.

For those who protect us.

acknowledgments

I was going to start this acknowledgment with the idiom “it takes a village,” but very soon I realized that, for my dissertation, it would be an egregious understatement: My dissertation took at least several villages distributed across the world, and this acknowledgment is my small gesture of gratitude.

First, to my advisors, Casey Fiesler and Jed Brubaker. I am grateful that they took a chance on me when I was applying to Ph.D. programs, supported my passion and interests along the way, and helped me turn my passion into tangible research products. I was always excited to discuss my work with them at any stage, knowing that they would have the magic to help me take it to the next level. Not only were they research advisors, but they were also my life mentors who guided many aspects of my Ph.D. adventure beyond research. I can only wish to emulate your kindness and rigor, and I hope that I have done you proud.

Also, thank you to my committee members — Brian Keegan, Amy Bruckman, and Eric Gilbert. It still feels unbelievable that I could have these rockstar scholars to support my dissertation, and their feedback has imporved my dissertation research significantly.

I am deeply fortunate to have shared my Ph.D. journey with an incredibly diverse group of colleagues. To Brianna Dym, Mikhaila Friske, Katie Gach, Anthony Pinter, and Morgan Klaus Scheuerman, who always so generously helped me brainstorm ideas and perfect my work. I want to especially acknowledge Morgan, with whom I got to collaborate on some of the craziest and most impactful projects. Thank you for being such amazing labmates, collaborators, and friends.

I am also grateful to the entire Identity Lab, the Internet Rules Lab, and the broader CU INFO research community for supporting me over the years and watching me do practice talks way too many times. My thanks to Jes Feuston, who joined CU at the tail end of my Ph.D. but offered me *so* much invaluable feedback. I'd especially like to acknowledge the amazing current and former administrative staff at INFO: Sarah Mandos, Amanda Robinson, Todd Amodeo, and Robby Rigby. They have been instrumental in my Ph.D. journey, and they deserve so much credit than they typically receive.

I had the privilege to work with some talented collaborators inside and outside CU: James Dykes, Benjamin Mako Hill, Charlie Kiene, Skyler Middler, Peipei Nie, Melanie Sidwell, Katta Spiel, and Kandrea Wade. None of my work would have been possible without them. I also learned much from academic friends who I did not have the opportunity to collaborate with, but helped me in ways they probably did not realize: Jeremy Birnholtz, Lindsay Blackwell, Stevie Chancellor, Eshwar Chandrasekharan, Bryan Dosono, Brent Hecht, Shagun Jhaver, Alex Leavitt, Kat Lo, Joseph Seering, Bryan Semaan, Kate Shores, Diyi Yang, and Amy Zhang.

I would like to thank researchers and mentors at Facebook and Yahoo!, especially Jess Bodford, Frank Bentley, Mark Handel, Frank Kanayet, Rushani Lyon, and Katie Quehl. I am also grateful to Facebook Research for supporting a significant portion of my dissertation research. In addition, my research participants brought my research to life by being willing to share their fascinating stories, and I am deeply grateful to them.

There is one person I am especially indebted to: My undergraduate advisor, Lana Yarosh. She was the one that first brought me onto my research career by allowing the undergraduate me to lead a research project. Thank you for pushing me to do what I did not think I could do.

There are also many people to thank outside the academy. First, to my family: Zongtao Jiang, Yan Cheng, Aiying Li, and my late grandfather Peihai Jiang. They have been tolerating me long before I started my Ph.D., and more importantly, trying and learning to support me in the way that I needed. In particular, my grandmother Aiying Li, who only received elementary school education, supported my early interest in computing when I was 13, as well as my wish to pursue education 6,000 miles away without any question. Also, my gratitude to my dearest friends over the years: Peng Liang, Hongze Liu, Yunzhi Liu, Chaofan Sun, Di Xie, and Rui Xue, for their unwavering support and many, many hourlong phone calls and video chats. Thank you for being my chosen family. Additionally, I want to thank Kyoto Animation, ufotable, White Fox, Sega, Square Enix, and Atlus for producing amazing anime and games that I desperately needed.

Finally, my deepest appreciation to Xiaozhe Zhang, who sacrificed so much during this Ph.D. journey and gave me more patience and love than I deserve. Thank you for being part of my life.

contents

tables

figures

1
introduction

Online communities have never been a utopia; they have been plagued with problems since the beginning of their time. For example, in Julian Dibbell's book My Tiny Life (1998), he describes "a rape in cyberspace" in which Mr. Bungle, a user of the text-based community LambdaMOO, "raped" other users by writing a program to force other users' programmable avatars to have virtual sex with him and with each other. This incident of online violence was unheard of to these early online community users, and LambdaMOO did not have any policy against cyberrape. Mr. Bungle's behavior not only outraged LambdaMOO users, but also instigated critical questions about the boundaries of acceptable and unacceptable behaviors online, and how LambdaMOO should be regulated.

While the cyberrape in LambdaMOO was eventually resolved by the administrator disabling Mr. Bungle's account, it was by no means the end of malicious behaviors in online communities. News is rife with stories of users of 4chan promoting hate speech and violence in the wake of a mass

shooting, or malicious people spreading misinformation aiming to prohibit civic participation. A recent survey from the Pew Research Center has shown that 41% of Americans have been harassed online, and 66% have witnessed harassment directed at others (Duggan, 2017). Terrorist groups are also widely using social media platforms for recruiting and spreading propaganda (Hossain, 2015). The range of harmful content has significantly increased since the time of Mr. Bungle, and it is clear that online communities today are still facing similar, but more complicated and challenging problems.

One solution to these malicious behaviors is content moderation. James Grimmelmann (2015) defines content moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." The "community" here can be big like the two-billion-user social media platform Facebook, which prides itself on being a "global community," or small like personal subreddits with a handful of people (Fiesler et al., 2018). However, they are all communities, and they all moderate—not only do platforms have complicated rules and policies that dictate what behaviors are not acceptable and hire professional content reviewers to sift through violating content, but many volunteer moderators also devote their time to combating these problems in their own, smaller communities on a daily basis.

There have been many success stories of moderating online communities, such as using automated moderation tools to automatically remove large amounts of problematic content (Jhaver, Birman, et al., 2019; Kiene et al., 2016), and setting positive examples to encourage similar behaviors in Twitch chat (Seering et al., 2017). On the platform level, prior research also shows that banning hate communities reduces the amount of overall hate speech (Chandrasekharan et al., 2017). Moreover, when the original hate community members migrate to new communities, these

new communities do not inherit their behavior from the old hate communities (Chandrasekharan et al., 2017).

Despite the hope and promise for content moderation to be the remedy of the internet's problems, it still faces many challenges, and they often arise from designing moderation solutions for one group of stakeholders—that is, people involved in and impacted by content moderation—but expecting them to work for everyone. One example of such stakeholders is communities with different modes of communication. While moderators and technologists have come up with largely successful ways and tools to deal with rule-breaking textual posts to the point that a prominent debate in content moderation is the philosophical tension between "free speech" and "safe space" (Gibson, 2019), communities with new technological affordances have cultivated unprecedented ways to break the most uncontroversial rules: Malicious users would cause disruption by producing loud, disturbing noise in voice chats (J. A. Jiang et al., 2019), or by "physically" sexually attacking people in social virtual reality (VR) (Blackwell et al., 2019), but there was nothing that advanced content moderation tools could do with them—there was no persistent content to moderate.

The scale of global social platforms also introduces stakeholders not being designed for. Facebook, for example, in 2019 had only 15,000 moderators globally to moderate content generated by over two billion users (Fick & Dave, 2019). Research has called to attention and proposed solutions for various problems caused by such scale, including delayed moderation (Lampe and Resnick, 2004), user circumvention of moderation (Chancellor, Pater, et al., 2016; Gerrard, 2018), and the emotional trauma of looking at intense and traumatic content as their day-to-day job (Roberts, 2019). While existing efforts have focused on the token "moderators" and "users" and the cat-and-mouse game between these two groups, developing community guidelines that cover abusive

behavior in the most nuanced details describable by English (J. A. Jiang et al., 2020), there is a large group of stakeholders being ignored—people in non-English-speaking regions speaking nondominant languages. Facebook's Community Standards, for example, were only translated into 41 out of the 111 languages that the platform supported as of 2019, which means people speaking the other 70 languages might not even know what the rules were (Fick & Dave, 2019; Wijeratne, 2020). Recognizing the global users that social platforms serve, Gillespie (2018a) noted that scaling content moderation is still an open and challenging question with today's incredibly complex online communities with millions of users of diverse backgrounds, purposes, and values.

As online communities increasingly play a key role in societal issues, the topic of content moderation is also gaining attention from the general public as well as the academic community. Critical questions have bee rasied about how we can address these challenges to make moderation more effective, many researchers have set out to study various communities, learn about their problems, and propose solutions. However, these studies often focus on one group of stakeholders —typically moderators or users—and the solutions that they offer often present themselves as widely applicable to content moderation in general. How these research insights differ for different groups of stakeholders, and how the proposed solutions will play out in different kinds of communities with different contexts are still unknown.

Together, these challenges illustrate that current practices and insights about content moderation can become questionable when additional stakeholders are introduced. While researchers and technologists have been trying to take more stakeholders into account, their effort is largely centered around one dimension—that of different functional roles in content moderation, typically moderators, users, and sometimes decision makers of social platforms. However, problems also exist

but often go unnoticed on other dimensions along which classifications of stakeholders occur, like the dimension of community technology or of geographical regions that the examples above demonstrate.

This dissertation takes a multi-stakeholder perspective to explore the challenges of content moderation, by considering multiple *dimensions* where stakeholders can be categorized and tensions that arise from them. My dissertation first closely investigates the two overlooked dimensions mentioned above—technological infrastructure and geographical region—as case studies, then holistically examines existing moderation research with an eye toward stakeholder tensions with the following themes and research questions:

- **Moderating Different Technologies.** How do different community technological infrastructures impact content moderation? How can moderators and designers cater to the unique needs of individual communities with different technologies?
- **Moderating Different People.** How does content moderation change for people from different parts of the world? How should platforms moderate content from people with different cultures and backgrounds?
- **Multi-stakeholder Tensions in Content Moderation.** What are the existing moderation strategies, problems, and potential solutions that researchers have identified? What do they reveal about multi-stakeholder content moderation when they are taken together?

In my dissertation research, I use both qualitative and quantitative methods to answer these questions. Blending post-positivist and interpretivist perspectives, my research approach combines deep investigations of people's lived experiences, as well as large-scale rigorous measurements of

STUDY SUMMARY	PUBLICATION	CHAPTER
Semi-structured interviews with Discord moderators about their experiences and challenges in moderating voice chat	Jiang, J.A., Kiene, C., Middler, S., Brubaker, J.R., and Fiesler, C. Moderation Challenges in Voice-based Communities on Discord. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 55 (November 2019).	Chapter 3
Content Analysis of community guidelines on 11 social media platforms	Jiang, J.A., Middler, S., Brubaker, J.R., and Fiesler, C. Characterizing Community Guidelines on Social Media Platforms. *CSCW 2020 Companion.*	Chapter 4 (Phase 1)
Survey of 2000+ global Facebook users in 10 geographic regions about their perceptions of abusive behavior	In progress	Chapter 4 (Phase 2)
Systematic literature review of empirical content moderation research	Submitted to CSCW 2021	Chapter 5

Table 1-1. Overview of Dissertation Research.

people's perceptions and opinions. My research also situates these insights more broadly within prior research, and considers what they reveal about content moderation as a whole. **Error! Reference source not found.** shows an overview of the research studies presented in this dissertation.

1.1 Contributions

At a high level, this dissertation contributes a multi-stakeholder perspective to content moderation research. Social computing research primarily studies socio-technical systems, which inevitably involve different groups of people with different needs and values. My research argues for an approach that positions different stakeholders and the trade-offs in meeting their needs at the center of designing socio-technical systems, instead of an one-size-fits-all approach that selectively, sometimes arbitrarily, dismisses some stakeholders. Content moderation, by being the arbiter of

what is "good" and "bad" in socio-technical systems, serves as an illuminating example of multi-stakeholder tensions.

At a more granular level, this dissertation offers contributions in the following areas:

- **Moderating Different Technologies.** Through a case study of moderation in voice-based communities, this dissertation reveals how new technological affordances introduce new complexities and challenges to content moderation, and how moderators reacted to them. This dissertation also contributes to larger questions around how moderation and regulation should advance in a time when technologies are quickly evolving and offering new ways to use and abuse, and argues that a single set of moderation strategies and tools should not be used unconditionally in different communities that may have different infrastructures and needs.
- **Moderating Different People.** Through a study that quantitatively measure global users' perceptions of abusive behavior online, this dissertation provides actionable guidelines for platforms to effectively moderate global content with limited moderation capacity. By identifying cultural and regional differences in the perceptions of violations, this work also raises critical questions around the efficacy of having a singular platform policy to govern global users with different values.
- **Conceptualizing Content Moderation as Trade-offs.** Through a systematic review of empirical content moderation research, this dissertation presents a trade-off-centered framework of content moderation by examining individual insights from focused case studies within the broader literature. It also demonstrates that two definitional components of

content moderation—facilitating cooperation and preventing abuse—are in tension in practice. In other words, trade-offs define content moderation.

- **A Multi-stakeholder Perspective of Improving Online Content Moderation.** Taking these studies together, this dissertation makes the argument that a core challenge of content moderation is how to balance stakeholders' needs and values. To this end, this dissertation calls for a multi-stakeholder approach to studying and designing content moderation, and advocates putting stakeholders' differing needs and values at the forefront when considering moderation problems.

1.2 Organization of the Dissertation

This dissertation is organized into six chapters. Following the current introduction chapter, Chapter 2 introduces the theoretical underpinnings behind online content moderation and online regulation in general, as well as an overview of relevant research in commercial content moderation and volunteer content moderation.

Chapter 3 describes the challenges caused by ignoring stakeholder needs through an empirical study. It introduces voice-based online communities as a group of stakeholders that render current state-of-the-art moderation approaches insufficient in the current content moderation landscape where most communities are text-based, and details problems that moderators faced due to a lack of moderation tools tailored for their specific technological infrastrucuture.

Chapter 4 focuses on the people side of content moderation, and describes an empirical study that considers users on a global scale beyond the typical Western perspective. It describes major types of abusive behavior across major social media platforms and differing perceptions of

them from users across the world, revealing the severe limitations of using a single set of rules to govern global users.

Chapter 5 moves from empirical research to a meta-study of content moderation research literature. It juxtaposes prior research findings and describes the pervasive trade-offs behind them, and concludes with a framework of content moderation that puts trade-offs at the center.

Finally, in my concluding chapter, Chapter 6, I reflect on the research in the preceding chapters and discuss recommendation and implications resulted from it. I close with a vision for a multi-stakeholder future for online content moderation.

2
background

In this chapter, I describe the background and prior literature relevant to my dissertation. I start by laying a theoretical foundation and introducing the idea of regulation and moderation. Then, I discuss two major categories of content moderation that cuts across the individual studies in this dissertation—commercial moderation and volunteer moderation—and highlight central problems in these areas and opportunities for this dissertation to address them. Finally, I briefly describe stakeholder theory and multi-stakeholder approaches in other governance structures.

2.1 Moderation and Regulation: A Theoretical Overview

In their book about online communities, Kraut and Resnick (2011) put forth effective regulation as one of the critical factors that make online communities successful. Moderation as arguably the most common way to enact regulation can facilitate civil discussions, resolve conflicts,

and ultimately make an online community enjoyable for its members. However, just like John Bercow, the "moderator" of the British Parliament, has more important responsibilities than shouting "order," moderation is more than a few people getting things back in order when problems happen. Tarleton Gillespie, in his book about content moderation, directly refutes the idea that moderation is peripheral and janitorial work like turning off the lights and sweeping the floors; instead, he argues that moderation is central and definitional to a platform's services and is "the commodity that a platform offers." (Gillespie, 2018a)

As mentioned above, moderation is the way to enact regulation in online communities, and Lawrence Lessig's "pathetic dot" model (2006) provides a way for us to understand regulation in general. Lessig points out that regulation comes from four sources: law, norms, architecture, and market. *Law* regulates by imposing legal sanctions on its violations; *norms* regulate through sanctions and constraints imposed by the community; *market* regulates by pricing and rewarding different actions; *architecture* regulates by its own features that constrain how people can behave. These four regulating forces are interdependent of each other, can support or undermine each other, and can make each other possible or impossible.

Lessig's four sources of regulation have some direct mappings in the context of online community moderation, described by Grimmelmann's (2015) taxonomy of the "verbs," or techniques, of moderation. Grimmelmann lists four different techniques of moderation: excluding, pricing, organizing, and norm-setting.

Excluding refers to keeping unwanted members out of the community, such as trolls and spammers. Excluding enacts regulation through architecture—it essentially removes access for certain members—and it can be very effective in targeting unwanted behavior. The action of

excluding can also range from being highly precise like individual-level bans to being extremely crude such as making an entire community inaccessible to newcomers. Excluding can also happen to entire communities, such as Reddit's ban of several hate speech communities in 2015 (Chandrasekharan et al., 2017).

Pricing enacts regulation by raising the cost of participating in a community. While some pricing structures inhibit participation through the source of market, such as World of Warcraft whose subscription costs between $12.99 to $14.99 per month, many other pricing structures are not monetary. The most recent example is Reddit's decision to quarantine r/the_donald, the subreddit dedicated to Donald J. Trump, the 45th President of the United States, due to its repeated offense of Reddit's content policy (Robertson, 2019). By removing the subreddit from search results and recommendations, and explicitly asking people to opt-in before accessing the subreddit, the quarantine is asking people to pay with their effort, time, and potentially mental well-being to find and access this community. The quarantine puts a monetary price on the community too—quarantined communities generate no revenue (*Quarantined Subreddits*, n.d.).

Organizing enacts regulation by shaping "the flow of content from authors to readers." Organizing consists of several techniques, including deletion, editing, annotation, synthesis, filtering, and formatting. While there are no concrete, detailed examples where organizing made a difference in an online community, a research study that I collaborated on found that 45.32% of subreddits had rules about formatting (Fiesler et al., 2018). Additionally, many subreddits require their members to properly annotate their posts ("flairs"), and the posts that are not correctly annotated will be removed. Subreddit moderators can also, and often, "pin" certain posts so they always appear at the top of the page. These organizing practices show that organizing enacts multiple sources of

regulation: pricing because members have to spend the labor to properly flair their posts; architecture because pinning posts changes the way content is displayed; and norms because pinned posts essentially show members examples of desirable behavior.

Norm-setting, according to Grimmelmann, is the "biggest challenge and most important mission" of moderation. As indicated by its name, it enacts regulation through norms. Norm-setting is a complex topic that merits more discussion than the other categories. I will first explain the types of norms that exist, then discuss how they play a role in moderating online communities.

Norms exist in two categories: descriptive norms and injunctive norms (Cialdini et al., 1991). *Descriptive norms* are the norms of "is"; they provide evidence of what will likely be desirable action, and what is "normal": if everyone is doing something, it must be a good thing to do. In online communities, descriptive norms often appear as implicit rules or agreed-upon best practices. For example, fan communities often have the norm of attributing credit to content creators (Fiesler & Bruckman, 2019). Chapter 3 describes many descriptive norms in voice-based online communities in the form of implicit rules, though they are not a product of general agreement—they are implicit because they are too difficult to articulate.

Compared to descriptive norms, *injunctive norms* are the norms of "ought" that motivate actions by promising rewards or punishments. They appear as clear, explicit rules in online communities that specify approved and disapproved behaviors. On a platform, they exist in different forms as terms of services, community guidelines, or self-created rules of subcommunities with different levels of governing power. I explain these different types of rules in the following sections, 2.2 and 2.3.

These "verbs" of moderation can also be carried out in different flavors, what Grimmelmann describes as the "adverbs." First, moderation actions can be taken automatically by a program, or manually by a human. The choice between automated and human moderation is a tradeoff between cost and quality: Automated moderation is cheaper than human in carrying out a large number of moderation tasks (though the initial programming may be expensive), but it inherently lacks the ability to understand subtle contexts and adaptability to new kinds of expressions and new ways to break rules, a limitation well documented by prior research (Gillespie, 2018a; Seering et al., 2019). Humans, on the other hand, often are more robust in making decisions than machines that are purely rule-bound. In general, more human attention means better but costlier moderation.

Second, moderation can also happen transparently or secretly. Transparent moderation means to make moderation decisions and the reasons behind these decisions explicit to the community, while secret moderation hides some or all of the details. While transparency increases the legitimacy of moderation, it does require the additional work of making these details public, and in cases of automated moderation, it is not always easy to explain why the program took certain actions, especially if the program involves notoriously uninterpretable machine learning algorithms. However, secrecy is not necessarily all bad: being transparent may expose loopholes and ways to bypass regulations to malicious actors.

Third, moderation decisions can also be made *ex-ante*—using the infrastructure to allow or prohibit behaviors before they happen—or *ex-post*—fixing the problem and punishing the rule breaker after something has gone wrong. While *ex-ante* approaches can ensure the consistency of moderation, *ex-post* approaches allow prioritizing moderators' attention to where it is needed at the

cost of letting abuse happen by default. Most platforms today largely use *ex-post* moderation due to their massive scale, which I will expand on later.

Finally, moderation can be central by a single group of moderators, or distributed by the regular users. Central moderation is more efficient and more consistent, offering a single checkpoint for all content to be scrutinized. However, a single checkpoint also means a single point of failure, and when central moderation fails, distributed moderation's robustness becomes prominent. Most platforms take a hybrid approach mixing these two: There are moderators who are responsible for reviewing content, and users can also flag unwanted content that happened to get through the moderators.

These techniques and flavors of moderation do not address the law as a source of regulation, but this is not to say it does not exist in online communities—rather, the law widely exists at a higher level as the overruling power. Many platforms like Reddit, Twitter, and Discord all have terms that prohibit unlawful behaviors. There are also specific laws in existence that require the moderation of certain types of content. For example, in the United States there are strict regulations against child pornography and sexual exploitation (U.S. Department of Justice, 2015); An amendment to the Communication Decency Act, which removed liability protections for online platforms that knowingly assist, support, or facilitate sex trafficking, was also recently signed into law (Jackman, 2018). Beyond the United States, many countries in the European Union such as the United Kingdom and Denmark also legally prohibit hate speech against protected categories (The Public Order Act 1986, 1986; The Danish Penal Code, 1930).

Today, the specific moderation and regulation approaches discussed above are carried out in two major categories: commercial moderation and volunteer moderation, each operating at different scales and having different challenges, which I describe in the following sections.

2.2 Commercial Moderation

Social media platforms like Facebook, Twitter, and YouTube have become a central part of the social lives of billions of people across the world. While these platforms are not by themselves content producers, they are responsible for storing, organizing, and circulating a massive amount of content (Gillespie, 2018b; Roberts, 2019). Despite their claims of being impartial and their reluctance to regulate speech (Gillespie, 2010), many platforms are incentivized to moderate: not only to meet legal and policy requirements, but also to avoid losing users subject to malicious behaviors, to protect their corporate image, to placate advertisers who do not want to be associated with sketchy online communities, and to honor their own institutional ethics (Gillespie, 2018b; Klonick, 2018). Through a case study of Reddit's ban on hate communities, prior research has shown that these moderation efforts can be effective in improving online communities (Chandrasekharan et al., 2017).

Platforms often govern users with two sources of rules: terms of service and community guidelines. Terms of service serves a legal contract between the platform and the users that spells out each party's obligations, liabilities, and other disclaimers, often written in an attempt to avoid future litigation (Gillespie, 2018a). Therefore, it almost always contains legalese that is difficult for regular users to understand, and as a result, platform users often misinterpret terms of service provisions (Fiesler et al., 2016), which might subject them to legal ramifications unknowingly, as my research

in Chapter 3 shows. Jackson et al. (2014) have called for deeply integrating the role of policy into social computing research, pointing out that policy is deeply intertwined with design and practice.

Community guidelines, on the other hand, often use plain language that explains platforms' expectations of proper user behavior. Compared to terms of service, users are more likely to read community guidelines and understand them when they have the need to reference platform rules. Platforms impose community guidelines that are much more stringent than the law requires, prohibiting illegal behaviors such as posting child exploitation and human trafficking, but also policing upsetting content like harassment and commercial spam that could drive users away. Gillespie (2018b) argues that these rules are important not only because they contain abusive behaviors and set the norms of platforms, but they also help construct an ecosystem of governance—smaller platforms may look to larger ones for guidance, sometimes borrowing their languages directly, and the larger platforms may also adjust rules and policies in relation to each other. A recent example is Twitter's ban on political ads immediately following Facebook's reluctance of removing or fact-checking them (Conger, 2019). These rules, especially when they are different, reveal how platforms see themselves as the arbiters of cultural values. My dissertation research speaks to the complexity of rules by offering a comprehensive content analysis of the community standards on major social media platforms, and revealing differences in focus areas, norms, and values between platforms.

As platforms grow to have millions or even billions of users who produce massive amounts of content, this immense scale presents significant challenges to platform content moderation (Roberts, 2019). Del Harvey, the Vice President of Trust and Safety at Twitter, said in a 2014 TED talk:

> Say 99.999 percent of tweets pose no risk to anyone. There's no threat involved. … After you take out that 99.999 percent, that tiny percentage of tweets remaining works out to roughly 150,000 per month. **The sheer scale of what we're dealing with makes for a challenge.**

What are the implications of the challenge of scale? First, the vast amount of content eliminates the possibility of *ex-ante*, or proactive moderation, where moderators review the content before they can appear on the platform. Almost all platforms have to resort to *ex-post* moderation, which initially allows all content without review, and removes or filters questionable content after the fact (Gillespie, 2018b; Roberts, 2019). However, this approach does let malicious content have its intended impact, even if it only stays up for a short period of time. Second, moderating an immense amount of content also requires an immense amount of human workforce. In 2019, Facebook alone has hired 15,000 full-time moderators across the world to combat malicious content, but this number is only a drop in the ocean compared to the billions of people whose content they need to review (Newton, 2019). Some disturbing content like revenge porn (Vanian, 2017) therefore remains online for days, months, or even years simply due to the lack of moderation capacity (Gillespie 2017). Many users also try to circumvent moderation, which makes timely moderation even harder (Chancellor, Pater, et al., 2016; Gerrard, 2018).

As platforms become global, the massive amount of users from various cultures and backgrounds also bring a wide variety of norms and values, which also presents a challenge to scaling up human moderation labor (Gillespie 2018), especially in platforms with a single set of rules for users across the world. News is rife with controversial moderation decisions such as removing the Pulitzer Prize-winning Vietnam war photo featuring a naked girl (Scott & Isaac, 2016), or

exempting misinformation from politicians from removal or fact-checking. The debates and conversations around these topics reflect the clash between different backgrounds, purposes, and values. Recent work has argued for a "constitutional layer" in digital institutions to make changes that are sensitive to local contexts (Frey et al., 2019). Chapter 4 in this dissertation research speaks to potentially resolving this clash of values by examining how people from different cultures and regions perceive rule violations.

At the same time, the human moderators will need to look at every piece of reported content, whether violating or not, as their day-to-day, 24/7 job. Sarah Roberts (2019), in her work about commercial content moderators, pointed out that the factory-like routine of content moderation work has led many moderators to burn out. The constant viewing of disturbing and traumatizing content takes a heavy emotional toll on the moderators, who are reluctant to discuss their work with their family and friends to avoid burdening them. This emotional burden has attracted wide public attention through the news of content moderators suffering from PTSD-like symptoms and having to work in filthy and hostile environments (Newton, 2019).

2.3 Volunteer Moderation

While many platforms hire full-time content moderators to combat abuse for the whole platform according to the community guidelines, many platforms comprised of smaller communities, like Reddit, Discord, and Facebook Groups, also rely on volunteer moderators who manage their own communities. These volunteer moderators have extensive administrative power over their own communities, such as setting rules, removing content, and banning people (Seering et al., 2019). Their communities are also governed by their own rules (Fiesler et al., 2018), which are often more granular than the platform rules and pertain to their own community contexts—for

example, the first rule of the subreddit r/Otterable is "Your post must contain otters. Try not to post sad content about otters."

However, volunteer moderators' power does not extend to communities of which they are not moderators, nor do they have control over platform-level issues. The rules that they create also do not supersede platform-wide rules. These volunteer moderators are usually the initial founders of the community, or selected users who are most heavily involved in the community and invested in its success; the moderator selection process can often be formal, requiring written applications and multiple rounds of interviews (Matias, 2016b). Moderators are thus familiar with the norms and values of the community, and are well-positioned to set and enforce rules that regulate content and behavior (Diakopoulos & Naaman, 2011).

As mentioned before, these volunteer moderators are not employed by the platform, but users often see them as representatives of the platform, granting them the power to negotiate with the platform on behalf of their communities (Matias, 2016a). The social curation platform Reddit, for example, hosts more than a million smaller communities called subreddits, and through a mixed-methods analysis on 100,000 subreddits, a research study that I collaborated on has found that over half of them have their own rules covering a wide variety of issues including harassment, advertising, post formatting, and more (Fiesler et al., 2018). While these subreddits are all governed by platform-wide policies, only less than 4% of the subreddits in their dataset mentioned these policies. This result suggests that rulemaking in communities is highly contextual, and that community norms are more salient than platform norms. Through analyzing removed posts on Reddit, Chandrasekharan et al. (2018) found that these community norms existed on three levels: *macro*—norms universal to most parts of Reddit, *meso*—norms shared across certain groups of subreddit, and *micro*—norms

specific to individual subreddits. These three levels of norms show that even though the norms are not explicitly connected to platform-wide policies, some of them indeed apply to the entire Reddit.

In the HCI and social computing literature, there are many stories of volunteer moderation successfully addressing problems in communities, often with the help of technical tools offered by the platform, or built on the platform's technological infrastructure. For example, Seering et al. (2019) found that volunteer moderators on Twitch, a popular live streaming platform, relied on bots and chat mode settings to mitigate abusive behavior. Through a series of research, Kiene et al. (2019, 2016) showed that automated moderation tools were critical to handling massive community growth, and that moderators built their own moderation tools using the platform API in the absence of native moderation tools. In another line of research, Jhaver et al. (2019) argued for the importance of moderation transparency by showing removal explanations increased perceived fairness of moderation as well as future user engagement. Furthermore, they found explanations from automated tools were associated with higher retention of moderated users, and pointed to automated explanations as a promising way to deal with scale. Taken together, all of these examples demonstrate the key role that platform technology has in the success of volunteer moderation.

On the other hand, technology also bring many challenges to volunteer moderation. For example, Discord moderators struggled to enforce rules consistently due to the difficulty in archiving moderation decisions caused by the lack of logging functionality on the platform (Kiene et al., 2019). A study with 56 volunteer moderators on multiple platforms showed that compared to those on other platforms, Facebook group moderators had to additionally spend a significant amount of their time reviewing group joining requests, a technical feature absent on many other platforms (Seering et al., 2019). While the Reddit AutoModerator was effective in addressing clear-cut

problems at scale, moderators also complained that the tool has a steep learning curve that requires deep familiarity with regular expressions, and moderators who are less programming-savvy often ended up making the AutoModerator incorrectly mass-remove innocent posts (Jhaver, Birman, et al., 2019). Furthermore, the large number of rules needed in the AutoModerator also made debugging and fine-tuning difficult when false-positives occur.

Overall, it is evident that the platform technology has a significant impact on community moderators' work, for better or for worse, and communities using different technologies often have different challenges and needs. While the examples above consist of largely quality-of-life issues for moderators, Chapter 3 in this dissertation studies a case where a new community technology unexpectedly sabotaged volunteer moderation, prompting researchers and designers to more carefully consider technology as a dimension of stakeholders.

2.4 Multi-stakeholder Approaches in Governance

Other governance contexts have also embraced an approach to involve various stakeholders—namely, a multi-stakeholder approach. The concept of "stakeholders" traces back to early management literature, referring to multiple constituencies impacted by business entities like employees, suppliers, local communities, and others (Freeman, 1983; R. K. Mitchell et al., 1997), and stakeholder theory uses descriptive, instrumental, and normative approaches to describe and identify business practices and functions in a way that accounts for these different entities (Donaldson & Preston, 1995). While initially started in the field of business management, stakeholder theory directly informed the multi-stakeholder perspective that many global governance groups widely adopt, because such an approach is more inclusive, more legitimate, and ultimately more effective (Gleckman, 2018).

Even within the premise of internet governance, the multi-stakeholder approach is already common, forming highly important groups such as the Internet Assigned Numbers Authority (IANA) and the Internet Governance Forum (IGF). IGF, in particular, has trust and safety as one of the recurrent thematic tracks in its annual meetings, covering subtopics closely related to content moderation like spam, child safety, and hate speech (Internet Governance Forum, 2020). However, due to the nature of IGF as an international policy group, its consideration of stakeholders is largely limited to platforms and government agencies. It also only focuses on high-level global policy concerns, instead of the nuances of individual platforms and communities, as well as their day-to-day moderation practices.

Nevertheless, the widespread adoption of multi-stakeholder approaches shows the promise of taking a multi-stakeholder perspective in more granular moderation practices. This dissertation also extendd current muli-stakeholder approaches by arguing that limiting to any one concrete categorization of stakeholders may cause unexpected harm, and encounrages a consideration of multiple possible categorizations of stakeholders.

3
moderating different technologies

It is easy to assume that content moderation involves removing existing content, usually with clear indications of who the author is. However, as communities with new technologies and new communication media arise, these communities become a key group of stakeholders that challenges existing understandings of what it means to moderate content and interactions.

As an initial step of understanding multi-stakeholder perspectives of online content moderation, in this chapter I use Discord, a platform where voice chat is a dominant mode of communication instead of text, as an example of what the current conception of content moderation overlooks.[1] In traditional text-based communities, moderation work mostly involves moderators

[1] This work was published at CSCW 2019 (J. A. Jiang et al., 2019).

locating the problematic content, and then removing it and sometimes also punishing the poster. This is a process that many people would take for granted, but how does this process work in the context of real-time voice, a type of content that lasts for a short time without a persistent record? The moderation of ephemeral content raises a number of questions: How do moderators locate the content? How do moderators remove the content? How do moderators know who the speaker is? How do moderators know whether the rule breaking happened at all? In this chapter, I first describe new types of rules unique to voice and audio-based communities and new ways to break them. Then, I describe how moderators struggled to deal with these problems. Moderators tried to give warnings first but sometimes had to take actions based on hearsay and first impressions. To avoid making elaborate rules for every situation, moderators instead simply stated that they had the highest authority. I then detail how these problems point to moderators' shared struggle—acquiring evidence of rule breaking, and how moderators' evidence gathering strategies could fail in different scenarios. Finally, through the lens of Grimmelman's (2015) taxonomy of community moderation that focuses on techniques and tools, I argue that voice precludes moderators from using the tools that are commonplace in text-based communities, and fundamentally changes current assumptions and understandings about moderation.

3.1 Research Site: Discord

Discord[1] is a free cross-platform VoIP application that has over 200 million unique users as of December 2018. Communities on Discord are called "servers," a term I will use throughout this chapter to refer to these communities. Despite the term "server," they are not self-hosted but instead

[1] https://discordapp.com/

hosted centrally on Discord hardware. While originally designed for video gaming communities as a third-party voice-chatting tool during gameplay, Discord servers now cover a wide range of topics such as technology, art, and entertainment. Every user can create their own servers as they wish, even simply as general chat rooms with no specific purpose. The size of Discord servers ranges from small groups of friends with a handful of people, to massive communities with hundreds of thousands of members.

A server typically consists of separate virtual spaces called "channels," usually with their own purposes, such as announcements or topic-specific conversations. A channel can be either a text channel or a voice channel, but not both. Users can also directly contact other users they are friends or share servers with through direct messages with text, voice, and video capabilities. A screenshot of the Discord interface is shown in Figure 3-1.

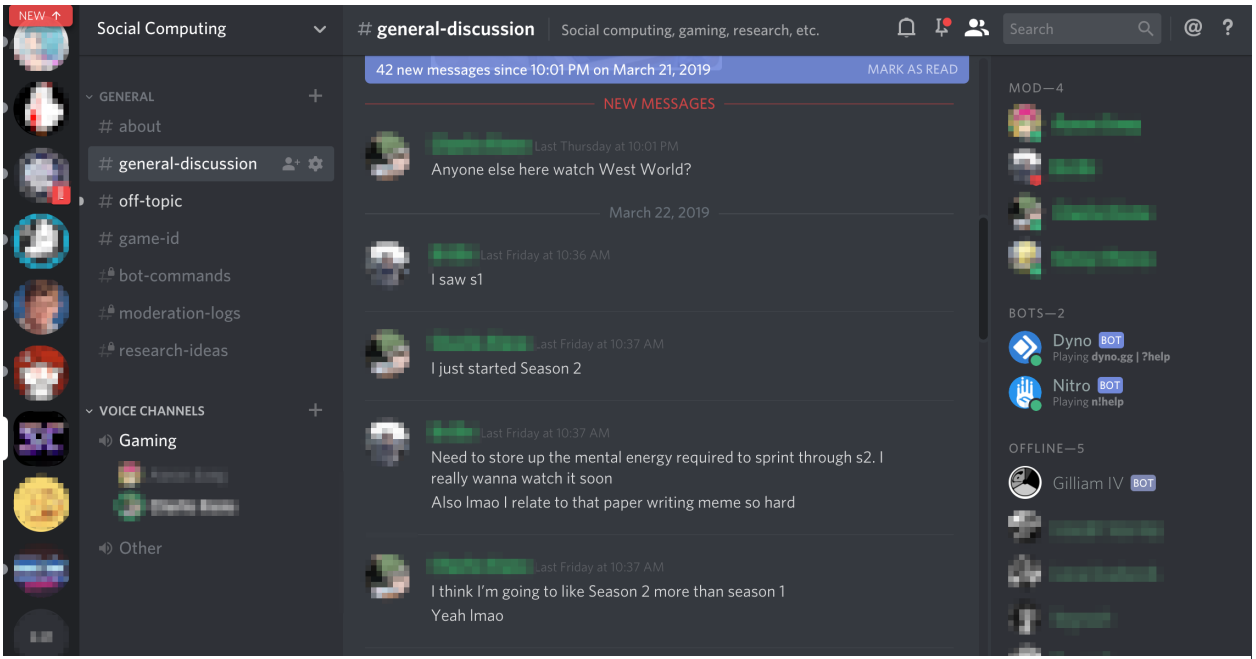

Figure 3-1. The Discord user interface. The far left sidebar lists all the Discord servers the user is a member of. The next side bar lists the text and voice channels of the Discord server the user is currently viewing. The middle area is for the scrolling text chat, and the right side bar lists the total users, categorized by their "role."

In voice channels, the only means of communication is real-time voice chat. Users can choose to have their mic open all the time, or push a button to talk depending on their settings. Discord does not provide ways to record or store voice chat, making them ephemeral. Users currently in a voice channel will appear in the user list of the channel, and will disappear when they exit the channel. A green circle around a user's profile picture indicates the user is currently speaking. Users can also mute themselves—make themselves not be heard—or deafen themselves—make themselves not hear everyone else and not be heard. Some Discord servers also have a special type of voice channel called "music queue," where a music bot plays from a member-curated playlist, and all other members are automatically muted.

Server creators can create different "roles" with custom names that grant users different permissions in the server, through which moderators gain their permissions as well. This role system allows for a hierarchy of moderation with lower-level moderators having less permissions, and higher-level ones having more. Depending on permissions granted to a given role, moderators can mute people, deafen people, or remove people from voice channels. Some moderators can also ban people from their servers, who will not be able to rejoin unless they are "unbanned."

While the forms of punishment provided by Discord are permanent by default, third-party applications called "bots" can be used to augment moderation by adding timers, making these actions temporary. Bots like MEE6[1], Dyno[2], and Tatsumaki[3] are well-regarded and widely used by

[1] https://mee6.xyz/
[2] https://dyno.gg/
[3] https://tatsumaki.xyz/

over a million servers to automate existing Discord features such as sending welcome messages and assigning roles. Besides improving existing moderator tools, many bots also provide additional functionalities for moderators, such as issuing people warnings that are permanently recorded in a moderator-only channel, and automatically removing content in text channels based on keywords or regular expressions. However, to the best of my knowledge, there are currently no bots with voice moderation capabilities.

3.2 Method

To understand moderators' experiences in moderating voice-based communities, my collaborator and I conducted in-depth, semi-structured interviews with moderators of Discord servers. Participants were recruited as part of a larger collaborative, IRB-approved project to investigate moderation in online communities. For this study we analyzed 25 interviews with moderators who identified as having experience in moderating Discord voice channels. We recruited participants by reaching out to moderators of open Discord servers. We also asked them to send the call for participation to other moderators, resulting in a snowball sample. The 25 participants came from 16 different Discord servers, with between 1 and 3 participants from each server. While the majority of the servers that we examined are large ones with more than one thousand members and may not be representative of smaller groups, I believe this over-representation is reasonable as formal moderation is less-needed in smaller communities (Kiene et al., 2019; Seering et al., 2019). The moderator participants provided us with a diversity of perspectives both across and within communities. Each participant was compensated US $20 for their time.

PARTICIPANT ID	AGE	GENDER	COUNTRY	SERVER TYPE	# MEMBERS
P01	18	Man	Croatia	Social Chatroom	164,257
P02	19	Man	US	Streamer	117,742
P03	19	Man	Russia		
P04	21	Man	US	Tech Support	233,762
P05	21	Man	India	Anime	130,924
P06	20	Man	US		
P07	18	Man	UK	Social Chatroom	57,319
P08	20	Woman	Malaysia		
P09	22	Man	US	NSFW	23,186
P10	23	Man	UK	Gaming	29,165
P11	39	Man	UK		
P12	23	Woman	US	Fandom	150
P13	24	Man	Australia	NSFW	55,239
P14	19	Man	US	Social Chatroom	77,512
P15	26	Man	US	Gaming	55,251
P16	24	Man	US		
P17	37	Man	US	Fiction Writing	1,137
P18	32	Woman	US		
P19	26	Woman	US	Gaming	3,246
P20	24	Woman	Netherlands	Gaming	24,542
P21	27	M	US		
P22	22	Woman	US		
P23	23	Man	Netherlands	Gaming	171,608
P24	24	Woman	UK	Gaming	63,001
P25	29	Man	US		

Table 3-1. Participant details of the interview study with Discord moderators. Numbers of server members are as of March 8, 2019.

Interviews ranged in length from 42 to 97 minutes, all of which were conducted over Discord voice chat. Participants' ages ranged from 18 to 39 (M = 24, SD = 5.43). Details about the participants, including age, gender, country of residence, and type and member count of the servers they moderate are presented in Table 3-1.

During the interviews, we asked participants to tell us specific stories about moderating voice channels, with follow up questions about how they found out about rule breaking, what specific actions they took, and what impact the incident had on the moderation team as well as the community. We also asked them to consider hypothetical scenarios, such as what participants would do if the rule breakers tried to evade punishment. Participants detailed a variety of moderation experiences that ranged in scale and in complexity. We also asked participants about the challenges of moderating voice channels, the behind-the-scenes deliberations of their moderator teams, and their feelings toward decisions they had made or situations they had encountered. Prior to analysis, all interviews were transcribed, anonymized, and assigned the participant IDs presented here.

With collaborators, I performed a thematic analysis of the interview transcripts (Braun & Clarke, 2006). I initially engaged in one round of independent coding, using an inductive open coding schema. All researchers on this project then discussed preliminary emerging code groups such as "catch in the act," or "enter the voice channel to confirm." Two more rounds of iterative coding helped me combine similar groups to create higher order categories such as "moderation challenges." I used these categories to produce a set of descriptive theme memos (Saldaña, 2009) that described each category with grounding in the interview data. All researchers on this project discussed the memos regularly to reveal the relationships between the categories and finally clarified the themes, which resulted in the three main findings I discuss below.

In describing the findings of this study, I start by characterizing new types of rules and new ways to break these rules in voice channels, then compare them to common types of rule violations in text communication. I then discuss the actions that moderators take to address these rule violations. Finally, I address the biggest challenge of rule enforcement in voice—acquiring evidence—by discussing moderators' strategies to gather evidence and how they often fail.

3.3 Rules in Voice and How People Break Them

Formal rules on Discord exist at the platform level in the form of Terms of Service and Community Guidelines, as well as at a community level in the form of custom rules set by the individual Discord servers. All the servers in my study had at least some explicit rules that were listed in specialized text channels, as well as implicit rules that were not written down but were nevertheless enforced by moderators. Though there were likely also emergent social norms in these communities, and rules that may have started out as norms, I spoke to moderators about the rules that they actively enforced, whether explicit or implicit, as opposed to norms enforced by the community itself. While there were many rules in the servers I examined, here I only focus on those with elements unique to voice.

Explicit Rules

Servers that I discussed with moderators had different explicit rules that governed both text and voice channels, such as "no advertising" or "English only," but all 16 of them had a rule against slurs and hate speech. I choose to take a deep dive on the rule of slurs and hate speech because it is the rule that most participants talked to us about, and presented challenges unique to voice.

Slurs and hate speech can make a community an unwelcoming and unsafe space for its members, and therefore many communities have rules against them (Fiesler et al., 2018). Just like in many text-based communities, slurs and hate speech are explicitly prohibited in voice channels, and are a major problem that moderators have to face. All participants told us that racial and homophobic slurs existed widely in their servers, both text and voice channels. In P08's server, racial slurs in voice channels faced an even harsher punishment than in text channels:

> Racial slurs in the [text] chat and VC [voice chat] are different. If you say it in the [text] chat, you get a four-hour mute depending on the severity, and in the VC, you get an instant ban because it's more … you know, saying it, rather than typing it, is much worse. (P08)

Racial slurs can be more offensive when spoken in smaller groups. Voice channels usually have 5 to 25 people participating at a time, which is much less than in text channels that typically have hundreds of active members. A potential consequence of the limited number of participants is that slurs in voice channels may feel more targeted and personal.

While slurs were not allowed in any of the servers, how moderators determined the threshold for what counted as a slur varied. For example, P03 treated the slur "n---er"and all of its intonations with a heavy hand:

> Like if you were saying, the "N" and then "word," then it's fine because it's not saying the slur itself. Any workaround … is not allowed. "N---a" or "n---a"—that's not allowed. Because you're still saying the slurs. Just rephrasing it. (P03)

Many moderators were cognizant of the different intonations a slur can have, and still decided to uniformly ban them—the only exception was the indirect reference "n-word." At the same time, while P06 agreed that different varieties of racial slurs were still racial slurs, he also took context and intention into account in his moderation, and the intonations did matter in his decisions:

> I think there's a difference between saying "What's good my n---a" and "I think all n---ers should go back to Africa." There's a pretty clear difference. So in that sense, you know, they're both racial slurs technically. And so by our rules … the punishment would be the same for both. But you know, again, it's kind of like a case-by-case thing. (P06)

While the intonations can still be expressed in text, P06's quote suggests that voice introduces more nuances to moderation, and that what technically counts as racial slurs by explicit rules may still receive different treatment. In text-based communities, moderation of content and intonations can be automated by using a list of keywords or regular expressions (e.g., Chancellor et al., 2016). But in voice communication where moderation cannot be as easily automated, and moderators have to hear everything for themselves, their personal judgments play a more important role. Having to moderate in a case-by-case manner also means more work for moderators.

Implicit Rules

While slurs and hate speech were explicitly against the rules in the Discord servers I examined, I also heard about behaviors that were not written in the rules, but that moderators still discouraged or prohibited. While moderators technically had the option to detail these in their explicit rules, these behaviors were complex, nuanced, and difficult to articulate. Below, I focus on

three main types of these behaviors that are unique to voice, and describe them as implicit rules in the servers: disruptive noise, music queue disruptions, and raids.

Disruptive Noise. Disruptive noise involves intentionally creating a loud or obnoxious sound in voice channels to irritate other people and disrupt conversations. According to many moderators, disruptive noise is a common rule violation in voice channels. One moderator, P14, said that their typical day involves "muting at least one person" in response to this kind of behavior. Disruptive noise points to several implicit rules that are not important in text-based communities, but stand out in voice. One of these rules is that one should not speak too loudly in a group conversation:

> I've had to step in because someone's told me "Oh there's a kid literally screaming down in the #underbelly" … So I'll hop in, and of course the kid will be screaming and he'll be muted. (P16)

The rule of not speaking too loudly shows a key difference between voice and text communication: voice is a limited-capacity communication channel. While typing in all caps does not affect other members' ability to type and be seen, speaking in high volume in a group voice channel takes up all capacity in the channel, effectively silencing others. Even though text spamming—posting lots of content repeatedly and quickly—also takes up channel capacity in a similar way, it is preventable by limiting the number of messages one can post in a given period of time (e.g., "slow mode" in Discord). Loud screaming, on the other hand, is not actively preventable on Discord unless a moderator steps in. Prior work has shown that people are already struggling with how to appropriately interject in group conversations (Isaacs & Tang, 1994) with their assumed

sequential nature (Clark & Brennan, 1991). The rule violation here is even more severe because it completely ignores turn-taking in conversations and forces the conversation to be about one person.

However, volume by itself was not the golden rule of determining whether someone is creating noise. As P14 suggested, hardware conditions also had an impact on someone's speaking volume:

> [Disruptive noise is] typically anything that would be considered "too loud." However, if someone has a sensitive mic, we typically let them know of this and make them aware how loud it is before taking action. (P14)

P14 further told us how he differentiated between intentional noise making and simply having a sensitive mic:

> When someone joins a VC [voice channel] for the sole purpose of "ear raping"[1], they will typically do so immediately; someone who joins a VC to use it for its intended purpose will typically start off with a greeting or some sort of intro. I try to be aware that not everyone is capable of purchasing the best microphones, so I try to take that into consideration. (P14)

While it may be possible to develop an automated program that detects disruptive noise by decibel value, P14's quotes show that in addition to presented behaviors, intention also matters. If intention is important for determining rule violations, automated tools clearly have limitations.

[1] Though I intentionally excluded it from all quotes but this one, I would like to note that a number of participants used the term "ear raping" to refer to creating disruptive noises. I hope that future work examines the use of the term in more depth, including the motivations behind its use and in what ways it may be more harmful to individuals or communities on the site.

Furthermore, the fact that the way someone initially joins a voice channel is important suggests a challenge in moderation: A moderator has to be present when someone joins, which is not a scalable solution when voice channels are always open and a moderator is always required. This constant presence is also not a reasonable request for volunteer moderators who are only contributing during their free time.

In addition to the appropriateness of volume, disruptive noise as a type of rule violation also points to the appropriateness of content in group conversation:

> There is one time I had my Discord on speaker and I was just talking to a group of friends. … [Some random people] joined and they started playing loud porn. So my brother was in the house … and he heard the porn blasting out of my speakers and he was like, "Yo dude, why are you listening to that at full blast on speaker?" (P06)

People tend to dislike unexpected auto-play audio on computers as it can cause physical discomfort and can be socially awkward in public (Ackerman et al., 1997; Chen, 2018). The inappropriateness of the content here only exacerbated the situation. Furthermore, people not directly part of the voice channel were also affected, which demonstrates how moderation work can potentially have a negative impact on moderators' lives outside the servers they moderate. P06 told us that the experience was "a little bit awkward," but it is likely that the same situation can happen in other contexts, such as in a workplace, with more serious consequences.

I also heard a similar story of disruptive noise where the sound itself was not irritating, but rather the way the sound was played:

> We'd have this weird phenomenon where there was a handful of people who would join voice for 5 seconds, leave for 2 minutes, then come back and join voice for 5 seconds, leave. So while you're playing all you would hear was "boop" "do-doop" [Discord notification sounds]—people just joining and leaving rapidly—and it was just so infuriating. (P10)

Just like in P06's example above, the unexpectedness also stands out in this example—people would not expect Discord features to be used in annoying ways. Furthermore, while it is possible to turn off these notification sounds, it is not feasible to do so in the middle of a game. There is also no way to ping a moderator without disengaging with the ongoing game, as in the case in P10's quote. This example also points to a difference between text and voice—in text channels, constant interruptions are accepted and somewhat expected. But in voice channels, the flow is much more important and even the slightest interruption can be disruptive.

Music Queue Disruption. Across my interviews, I heard stories of rule violations not only in conversational voice channels, but also in music queues—voice channels that automatically play from crowdsourced playlists. In music queues, members are automatically muted, but they can append new music to the shared playlist, and skip the music that is currently playing if the channel setting allows.

A rule that music queues shared with conversational voice channels was not to be too loud:

> Literally the most recent thing that happened. … Someone put [something] in the music queue and it was for like two hours of just extremely loud music. (P04)

P04's quote shows that there was another type of rule violation other than being too loud: the music was also long. According to P04, checking whether people put "hour-long shitposting tracks that nobody wants to listen to" in the music queue was his everyday job. Because music queues are necessarily sequential, a turn-taking rule becomes important: people agree to occupy the channel completely for an appropriate, limited amount of time. Compared to disruptive noise, the problem with playing long music is not that it takes up the whole channel—it is in fact expected—but is that it essentially rids other members of their turns by taking up the channel for a long time.

Another form of rule violation that stops other people from participating is to skip their music:

> I actually got mad at the user. … What they were doing was they were constantly taking our jukebox and turning off other people's stuff. Like someone would play something and then they would skip it. (P21)

This example, together with the previous example, emphasizes the importance of turn-taking in music queues. While turn-taking is not a problem in text because of its threading structure, it becomes a major problem in music queues that resemble town hall meetings, where everyone gets an opportunity to hold the mic—playing long music is like someone who will not stop talking, and skipping music is like forcing the person with the mic to stop talking.

Raids. In addition to rule violations by individuals, I also heard stories of organized rule violations that moderators called "raids." Raids are organized voice channel disruptions that involve multiple users that can be human or bots. P13, for example, experienced a raid that was similar to P06's story, but on a larger scale:

> There was one a few months ago where they were spamming porn over their mics and they all had profile pictures of the same girl in a pornographic pose. And there were maybe like 15 of them in the same voice chat. (P02)

While an individual playing pornographic audio is irritating by itself, one can only expect 15 people doing so would cause even more discomfort. While raiding violates the similar rules in voice that I mentioned above, it is considerably more difficult for a moderator to manage. In the case of an individual, a moderator only needs to take action on that person, but in the case of a raid, a moderator needs to act on all the people involved. Across my interviews, I heard stories of raids that involved up to thousands of bot accounts, but moderators could only manage the accounts one by one—there is currently no way to manage multiple accounts all at once. This restriction means that not only do the moderators have to take on a significant amount of work managing raids, but also there is no way to prevent the other raiders from evading once they see one of them is punished.

Fortunately, while managing raids could be difficult, recognizing raids was relatively easy. As P02's quote suggested, the raiders all had similar profile pictures, which often became a clear signal for raid detection:

> [W]hen I see multiple people with the same or similar names rapidly join a VC, that's a warning sign for me. (P14)

Moderators told us that when they saw people with similar profile pictures or names, they would join the voice channel to confirm if they were part of a raid. However, I also heard from some moderators that they saw these cues as definitive signals of raids, and punished these people without confirming. Moderators told us that there were no unfair punishments because they believed if there

had been any, these people would have appealed. This moderation approach can be potentially problematic in cases where members use similar usernames as part of a community in-joke, which I also heard in the interviews. While moderators' tolerance of false positives—incorrectly identifying someone as a raider—may be a reasonable attempt to reduce their workload, it also suggests that they could have unknowingly driven these members away for good.

3.4 Moderation Practices

To punish the rule breakers, moderators used tools provided by Discord, including muting, deafening, and banning for more serious offenses, which third-party bots enhanced by setting a timer on them, making these punishment temporary. While I found that the correspondence between rule violations and types of punishment was highly contextual to individual servers, there were still some common moderation approaches that the moderators took. Specifically, moderators tried to warn first before taking more severe actions, but sometimes they also took these actions merely based on hearsay or first impressions. Even though these signals were unreliable, moderators did not come up with specific rules that described acceptable and unacceptable behaviors, but developed catch-all rules that dictated their authorities instead.

Warn Before Punishment

Across my interviews, a common approach to moderate voice channels that I heard from 23 of 25 moderators was to warn first before giving out real punishment:

> We would unmute our mics, start talking casually to people and then we would just figure out if they are breaking the rules or not, warn them verbally in the

> voice channel because we do roll with things in voice channel. … We'd always do verbal warning before we do any other actions. (P07)

P07's quote points to a key difference between voice moderation and text moderation. In the case of text, moderators would not have to "figure out" if someone broke the rules—all the conversations are recorded and visible to moderators. However, because Discord does not record voice channels, there is no way for moderators to unambiguously determine rule violations. "Rolling with things" allowed moderators to learn the contexts of alleged rule violations while not upsetting community members, and giving warning first let moderators still enforce rules to some extent but not have to risk wrongly punishing someone.

According to the moderators, warning was not only a way to conservatively enforce rules, but also a lesson for the community:

> We have this weird thing where our moderation is kind of public in that, if we warn someone not to do that, we try to always do that publicly so that other people can say, "Hey, okay, that's an example of what not to do." (P19)

The public moderation that P19 mentioned suggests community members learn rules through examples, which prior research shows can be effective in discouraging undesired behaviors (Seering et al., 2017). While example-setting certainly has been successful in moderating text-based communities, explicitly showing what the rules are may be more important in voice-based communities because the rules might be ambiguous or unfamiliar, especially to newcomers who are more familiar with text-based communities.

Punishment Based on Hearsay and First Impressions

While in most cases moderators tried to give warnings first, they sometimes also punished rule breakers directly. For example, P11 told us about a time the moderation team banned a member only based on a member report, without extra deliberation:

> Someone complained that a user was harassing them in voice chat and just shouting profanities and racism down the mic, so we just banned the user and we didn't hear anything, and they didn't appeal it. … We kind of took it at face value. It's very hard to get evidence from a voice chat unless you're recording. (P11)

Here, P11's moderation team assumed the punishment was fair only because the person punished did not push back, which may be an acceptable solution without adequate evidence. However, punishment based on hearsay does risk wrongly punishing someone in the case of false reporting. In such a case, the person punished by mistake possibly would not have appealed anyway when they were frustrated by undeserved punishment and left the community (Lampe & Johnston, 2005).

Moderators sometimes also took actions based on their own understanding of the person involved. P21, for example, told me that "first impressions matter, especially in voice chat," to the extent that they justified the most severe form of punishment:

> So if a user has just joined … and immediately shows this kind of behavior, such as the derogatory terms, trolling, sexist behavior … they'll just get banned. We won't warn them. We'll ban you because you've made it quite apparent that you

> had no interest in reading the rules and you're not all that great of a person, to say the least. (P21)

P21's quote shows a stark contrast with the "warning first" approach of moderation, which he also took in his community. However, the importance of first impressions in voice suggests that a person's intention and character may be more salient than in text, and they can be used in moderation decisions.

Catch-all Rules

With moderators sometimes having to use unreliable cues like first impres-sions in moderation, one question arises: Couldn't they develop rules that unambiguously describe what behaviors were acceptable or not? Moderators gave their answers to this question:

> There are times where you can't catch everything with a rule set. So I think that's more the reason why we don't have an on paper set of rules because we are such a large server. We would have to have a million different rules to help cover everything. (P04)

Across the Discord servers I examined, only two had guidelines about acceptable behaviors specific to voice channels. P04's quote shows that the lack of guidelines comes from the practical standpoint that it is simply impossible to list every single type of behavior, which suggests that the variety of (mis)behaviors possible in voice is much more diverse than in text. However, this is not only a problem of quantity—the level of nuance involved in the rule violations themselves shows that the line between acceptable and unacceptable behaviors is also difficult to articulate.

The solution that servers used to solve this problem, including the two that had guidelines in voice channels, was to create catch-all rules that were different varieties of the same core idea: Moderators have the utmost authority.

> [If someone does something unacceptable] then they are well within their rights to turn around and say, "Well that isn't in the rules." And then it's just a nice get-out clause for the moderator to be able to turn around and say, "Well look, whatever a moderator says goes." (P11)

This umbrella rule did not only free the moderators from having to create new rules, but also reduced the potential work of arguing with members. As moderators had greater authority, they also took greater responsibility in making the right decisions. But as I heard from the moderators, the lack of evidence in voice became an obstacle. In the next section, I describe the ways moderator gathered evidence of rule violations and their unique challenges.

3.5 Acquiring Evidence

Moderators told us about their various practices and actions upon rule violations, but a prerequisite of any action is that moderators had to make sure that a rule was violated. Acquiring such evidence in ephemeral voice channels, however, was a major challenge:

> Voice channels just basically can't be moderated. … The thing is in voice, there's no record of it. So unless you actually heard it yourself, there's no way to know if they really said it.(P19)

Gathering evidence was such a challenge that moderating voice channels was almost impossible to P19, and this difficulty points directly to a fundamental difference between voice and

text. In text, it is common knowledge that every conversation is automatically recorded. Moderators would not even have to consider acquiring evidence because everything is persistent. Even for ephemeral text chat applications like Yik Yak, where community moderation mechanisms such as upvoting and downvoting are possible (Schlesinger et al., 2017), the text would still have to be persistent for a short period of time. However, this common assumption about text-based moderation breaks down completely in voice because of the ephemerality of real-time voice conversation.

Moderators came up with different strategies and workarounds to tackle the problem of obtaining evidence in an ephemeral environment. In the rest of this section, I describe three main types of strategies that I heard from moderators: (1) Entering voice channels when the rule violation was happening; (2) asking witnesses to confirm someone was breaking the rules; and (3) recording voice channels.

Entering Voice Channels to Confirm Rule Breaking

As P19 said, the most reliable way for moderators to acquire evidence was to hear it for themselves. However, constant presence in voice channels is not a reasonable request for volunteer moderators. Therefore, moderators took one step back—entering voice channels when they realized someone might be breaking the rules:

> When we do get a report, then if someone is available and it's going on right then and there, then we'll hop on and listen and see what's going on and see if we can determine any rule violations right there off the bat. (P25)

While entering voice channels may be a solution, P25's quote does point out a important requirement: a moderator has to be online at the time of rule violation, and the rule violation has to be still ongoing when the moderator joins. This requirement is difficult to fulfill, considering the time a member would take to report to a moderator, the time the moderator takes to see the report, and the time the moderator takes to join the voice channel. This requirement also means that, for any violations that are instant and do not extend over a period of time, moderating with this method is nearly impossible.

Even if a moderator was present at the right time, it was still not guaranteed that the moderator could identify the rule breaker:

> There's also the problem of telling who's talking, because … if everyone's talking at once and some big fight is happening, it's kind of hard to tell who's talking. (P21)

Discord has no platform limit on the number of people a voice channel can have, and many mod-erators had moderated large voice channels with as many as 25 people. Voice channels, unlike text channels with threading structures, is inherently limited in information capacity. Therefore, it can be difficult to tell who is saying what when many people are talking simultaneously. Furthermore, there was also another problem with multiple people in the same voice channel—there was no way to stop people from evading:

> When they start seeing people get banned, they leave the voice chat so they don't get banned themselves and then it makes it harder for us to find them. (P02)

Tracking down rule breakers, according to P20, was often a "wild goose chase, only on steroids." The need to track down rule breakers speaks to the importance of not only acquiring evidence, but also the information contained in the evidence, which is also different between in text and in voice. In text, the text itself contains evidence in the form of rule-breaking content (e.g., racial slurs), but also the metadata—who the poster was, when it was posted, etc. When a moderator need to punish a person, this information connects the problematic content to the person, allowing the moderator to act without the person and the content being present at the same time. In real-time voice where content is ephemeral, however, this metadata does not exist. Even though the moderator has the evidence—they heard the content—there is no persistent record that associates the content to the rule breaker. Furthermore, identities in voice channels are also ephemeral: once the person leaves the voice channel, their profile picture disappears, and the information that connects the voice to the person no longer exists. Without this metadata, in text the moderator can at least delete the content, but in voice, the content is ephemeral to start with, and the moderator can neither moderate the content nor punish the person.

Relying on Witness Reports

With joining voice channels being ineffective in many cases, some moderators turned to witnesses' testimony as evidence:

> If the violation has already passed and people have split up, then we'll get a list of the users that were in that voice channel at the time and we'll send a message to them and just ask them, "Hey, you heard that there was something going on with this voice channel, can you give us more information about this?" (P25)

Moderators hoped to use witnesses' memories as a way to overcome the ephemerality of voice, but note that P25 here contacted a list of members instead of a single member. Without concrete evidence, the quality of witness reports became important to separate legitimate reports from hearsays and rumors, and one proxy for quality was the number of witnesses:

> We need witnesses like at least maybe two other people … that were there, that can confirm like, "Yes, this person said that." (P02)

P02's quote suggests a simple rule of thumb: the more witnesses there were, the more credible the report was. While moderators spoke of this as a generally successful approach, there were nevertheless problems with it. First, multiple witnesses would work only if their stories were consistent, but when they were not, moderators were likely to fall into the rabbit hole of seeking more witnesses and more stories. This could lead to the moderation workload no longer being comparable to the severity of the violation anymore. Second, even if the stories were consistent, moderators had no way to make sure the stories were true:

> You just have to be careful with the report, because some people can do false reporting, to do whatever to the other user. (P24)

P24's quote reveals an interesting possible scenario, where one member falsely accuses another member due to some personal vendetta. Requiring multiple witnesses may be able to mitigate this problem between dyads, but could facilitate a planned brigade against a community member, and in this case the member could not even appeal their case, again, due to the lack of concrete evidence. To mitigate the problem of false reporting, some moderators came up with a workaround—sending "spies" into the voice channel:

> We have these accounts that don't have the staff role, you know, you just send those accounts and if someone catches them they're able to get warned. … You're basically sending someone who is not a staff to catch that guy. (P05)

Three moderators told us that they asked trusted members to be "spies," or sent in undercover moderators into voice channels. This practice shows the complexity of moderators' evidence acquisition strategy, and also suggests a difference in the expectation of privacy between text and voice. In text, people assume that everything they write will be accessible by the moderators, but real-time voice chat breaks down this notion and suggests an increased amount of privacy—people cannot hear what someone is saying unless they decide to join the same voice channel and be seen. When there are accounts that do not truthfully represent their identities, people in the voice channel no longer know (1) who they are talking to, or (2) who can hear their conversation. One may think that people should not break the rules anyway and they will not be impacted if they do not break the rules, but this *de facto* surveillance system can create a chilling effect that discourages members from using voice overall.

Recording

Recording voice channels is probably the most straightforward way to address the problem of evidence by making ephermeral content persistent, and allows moderators to discover rule-breaking voice in a similar way in text. However, instead of going into the voice channels and recording directly, moderators took more concealed approaches. P21, for example, used a method that he called "incognito recording":

> There were six users in voice chat … harassing a female player … so I was alerted and I joined and they immediately clammed up. I left and ... I had a user [who

> was not a moderator] join the voice chat and I recorded through that user to listen to the entire conversation. (P21)

P21 used a combination of voice recording and "spy" users, because the rule violators would immediately stop when a moderator joins. This example further breaks down the notion of privacy in voice channel described in the last section: people no longer know whether their conversation is on the record, on a platform where they expect ephemerality. Prior research showed people had strong negative reactions toward instituting permanence in a previously unarchived space (Hudson & Bruckman, 2004), and we can only speculate that doing so secretly will even further discourage people from participating.

While recording might be a guaranteed way for moderators to get evidence, moderators told us recording only voice was not enough:

> We only really take MP4 files at the moment ... and what it will do is it will come up with a little snippet of what the voice channel actually looks like with all the people in it. And because whenever someone talks in voice channel, it has a green circle around their avatar. What we do is we fathom out who's talking at that moment saying whatever they're saying and then the action it needs. (P07)

For open servers, it is impossible for moderators to connect voices to members' identities—moderators cannot remember everyone's voice and cannot possibly have heard everyone's voice in a server with tens of thousands of people, not to mention there are new people joining all the time. Screen recordings, according to P07, mitigated this problem because they indicated who was talking at the time (in this case green circle around speakers' profile picture), and essentially achieved the

effect as if moderators were there hearing the conversation for themselves. However, one problem I identified with moderators entering the voice channels—they could not tell who was breaking the rules when multiple people were talking at the same time—still persists with this solution. Furthermore, screen recording also introduces more technical overhead because the resulted file size would be considerably larger than textual recording or even voice recording. The problem of screen recording resonates with Grimmelmann's (2015) discussion of the need to contain the cost of moderation within acceptable levels.

Finally, one moderator raised a problem with recording that extended to the legal space:

> Recently we had like one example of, somebody reporting somebody for recording them in the voice channel, which is like a can of worms because they were like, according to Turkish regulation, or whatever where this user is from, it's illegal to record somebody without two party consent. (P23)

Laws regarding recording conversations are not consistent around the world. In eleven states in the United States, as well as in some other countries such as Germany and Turkey, two-party consent— that is, all parties of the conversation must consent to the recording—is required ("Telephone Call Recording Laws," 2019). Sending the recording to other people is also against the law in countries like Denmark, which could potentially deem the behavior of the member who recorded for moderators illegal. However, currently there is no consent process for voice channel recording in either Discord's Terms of Service (other than a term that prohibits illegal activities), or the rules of the servers that I examined.

Overall, the findings revealed new kinds of behavior and norms that necessitated new rules in the context of voice communication, as well as new ways for people to break them. Moderators developed various practices to adapt to these changes, but they struggled to acquire evidence of rule breaking. While they developed tactics and workarounds to gather evidence, none of them were perfect, and all of them introduce problems that ranged from punishing an innocent community member to potentially breaking the law. These problems are unique to moderating voice, and challenge existing assumptions about moderation practices that are common in text.

3.6 Discussion

The stories I have heard revealed many unique moderation problems in voice that moderators had difficulty solving. Moderators were often unable to prove someone broke a rule, and when they did, they could only moderate in a post-hoc manner.

Participants' stories about moderation in voice channels challenge many understandings and assumptions about moderation based on text-only communities. In his analysis of text-based communities, Grimmelmann lists four moderation techniques, or "the basic actions moderators can take": excluding, pricing, organizing, and norm-setting. All of these techniques are available in text-based communities, but real-time voice communication rids moderators of many of them.

In open text-only communities, organizing seems to be the most common moderation approach based on prior work (e.g., Kiene et al., 2016; Seering et al., 2019): deleting posts, editing the content of posts, annotating posts (such as adding tags and flair), etc. However, these basic, commonly-used abilities completely cease to exist in real-time, ephemeral voice chat: there is currently no way to delete, edit, or annotate someone's voice as they speak. In open communities

where pricing is not an option, moderators of voice-based communities are left with only two options: excluding rule breakers from the community, or setting desired norms. However, neither of these is without its problems.

In text-only communities, strategies that exclude rule breakers, such as muting or banning them, are typically a last resort for severe cases since moderators are able to moderate the problematic content itself. As long as the case can be handled in terms of only content, actions on the person who posted the content are usually unnecessary. On the other hand, in voice channels, banning or muting someone is the only way to prevent problematic content from appearing, because that content is entirely dependent on the person's presence. In other words, banning content is equivalent to banning a person's presence in voice, either permanently or temporarily.

Norm-setting is also challenging in voice-based communities. According to Grimmelmann (2015), moderators can set norms directly by making rules, or indirectly by showing examples. However, my findings suggest that moderators avoid direct norm-setting due to the greater variety and complexity of ways to break rules in voice. Compared to text, it is less feasible for moderators to create a separate rule for every possible violation, not to mention some of which might be too nuanced for rules to articulate. While my findings show that moderators do engage in public example setting as a way to indirectly influence norms, it can take a long time for norms to emerge and moderators still have to take the more extreme action of excluding in order to meet their short term need to remove problematic content.

In addition to rendering these basic actions unusable, voice also restricts moderators in terms of the tools they can use. For example, moderators cannot auto-moderate voice because there is currently no equivalent to keyword-filtering. Instead, moderators can only react because there is no

way to preemptively prevent rule violations from happening in voice at the level of infrastructure. Compared to text, in which moderators have many tools at their disposal, in voice moderators are fighting with their bare hands, against more and harder problems. As articulated by Clark and Brennan (1991), the medium used can dramatically change the availability, cost, and effort required for communication techniques; my analysis reveals that these challenges due to change in medium apply to not just grounding, but also moderating communication.

Beyond the challenges of executing moderation when they know that a rule has been broken, there are additional challenges for moderators to even determine whether a rule has been broken. The participants emphasized a critical aspect of their practice that does not appear in prior work focusing on moderation tools and strategies (e.g., Grimmelmann, 2015; Kou & Nardi, 2013): the need for evidence. In text, there possibly is no such need—because until a moderator deletes it, text will always be there. On the other hand, moderators told us that evidence was a major problem in voice channels, due to the ephemerality of voice. While moderators developed strategies to address this issue, such as entering voice channels when receiving a report, or recording, these strategies were unreliable at best, and at worst, they risked breaking the law.

In sum, these findings show that overlooking community technological infrastructures could lead to serious challenges for stakeholders who employ technology for which existing moderation practices do not support. In the next chapter, I turn my attention to challenges risen in a different dimension of stakeholders—geographical regions.

4
moderating different people

In the previous chapter, I revealed the challenges of moderating communities with different technological infrastructures, and the potential consequences of failing to do so. However, in addition to different technologies, content moderation also faces the difficult task of managing different people. As social media platforms have grown into global scale, the people that they need to moderate have also become increasingly diverse. As a result, moderating these people, as well as the enormous amount of content they generate, becomes a significant challenge.

Take Facebook, one of the currently largest social media platforms, for example: Facebook has 2.7 billion monthly active users worldwide, and supports 111 languages. However, it only has 15,000 moderators (Fick & Dave, 2019) to regulate billions of users with one set of rules, the Facebook Community Standards (Facebook, n.d.). Moderators need to enforce the Community

Standards consistently throughout the world, but how do we ensure consistency when the moderators and the people they moderate come from different backgrounds, cultures, and values? For example, prior research has shown that the acceptability toward sexualized nudity varies between cultures (Smith, 1980), and news media have also shown that translation of community guidelines did not extend to all languages. Therefore, when a moderator in Hyderabad needs to review content created by someone in Austin as a result of global distributed moderation (Roberts, 2019), do they perceive the content and the governing rules consistently? Furthermore, as the moderation resources are limited compared to the extremely large amount of content that they need to handle, how do we prioritize the resources where they are most needed? And how does the prioritization change in different parts of the world?

Results from Chapter 3, as well as other research (e.g., Blackwell et al., 2019; Chancellor, Lin, et al., 2016), have suggested that severity can be a good heuristic to differentiate and prioritize different types of abusive behavior: Those that are more severe should be prioritized than those less severe. A deep understanding of severity will provide actionable guidelines for moderators to make more informed moderation decisions, for social media platforms to more effectively prioritize moderation capacity, and for regular users to better understand rules online as well the different levels of consequences of violating them. Therefore, in this chapter, I quantitatively investigate the severity of different types of abusive behavior through a multi-phase, large-scale study with people from ten different regions in the world.

The scale and complexity of study required that I conduct it in two phases. First, in order to understand the severity of abusive behavior, understanding the existing types of such behavior is a

prerequisite. Therefore, in Phase 1 of this research, I present a content analysis that resulted in a comprehensive taxonomy of abusive behavior across 11 social media platforms.

Having identified the types of abusive behavior, in Phase 2, I move on to understand the severity of these types of behavior. I first describe a novel method to measure participants' perceptions of the severity of these behavior types. In presenting the findings, I first discuss the perceived relationship between the severity of different types of abuse, specifically how the perceived severity grows as abusive behavior gets worse. I then describe the similarities and differences between perceptions in different geographical regions, how regions agree or disagree with different types of abusive behavior, and topics that are regionally sensitive.

4.1 Phase 1: Community Guidelines Content Analysis

Before I can investigate the severity of abusive behavior, a necessary first step is to identify what types of such behavior exist. Community guidelines are a promising source for descriptions of abusive behavior. Social media platforms use community guidelines to specify the abusive behavior that they prohibit. While documents such as Terms of Services and Privacy Policies also act as sources of regulation, they are usually legally binding, written in legalese that are difficult for an average user to understand (Fiesler et al., 2016). Community guidelines, on the other hand, are written in plain language, designed for ease of understanding for regular users. These community guidelines are often more granular and stringent on users' behavior than what the law requires, and while there are rarely legal consequences for violating community guidelines (except for illegal behavior such as posting child pornography), it can lead to rule-breakers receiving punishment such as restricted access or bans on their accounts (Gillespie, 2018a). Platform-wide community guidelines are also different from subcommunities' own rules (such as those of subreddits) as they

govern all subcommunities on the platform regardless of what each subcommunity's rules are. Different platforms' community guidelines also address different types of behavior (with some overlaps) in different granularities, as I will show in my findings below.

With a goal to collect community guidelines that are most comprehensive and to generate a taxonomy generalizable to all platforms, a research assistant and I chose the 15 social media platforms with the most monthly active users based on published statistics (Clement, 2019b), because platforms with more users are likely to also have more detailed rules. After excluding platforms that do not have published community standards or guidelines in English (WeChat, Qzone, Sina Weibo, and Douban), we generated the final list of social media platforms for analysis, shown in Table B-1 in Appendix B. We read the community guidelines on these social media platforms in November 2019, and independently coded for emergent types of behavior and then came together to adjudicate differences and iterate on codes. We did not find any new codes after coding for the rules on Facebook and YouTube, the two platforms that have the most extensive set of rules in our dataset. The final codebook revealed a total of 66 different types of rules across all platforms, also shown in Table B-1 in Appendix B.

Then, using the codebook, we both independently coded the rest of the platforms, and checked for interrater reliability. We achieved a Cohen's Kappa of at least .7 for every platform, which is higher than the threshold of "substantial agreement" (Landis & Koch, 1977). The researchers also discussed coding disagreements to ensure that they were due to reasonable subjective judgments and not systematic misunderstandings, and eventually came to an agreement on all codes. Based on this coding, I then analyzed patterns across platforms and types of abusive behavior that their rules covered.

Overall, I found significant variability in the coverage of infractions between the 11 social media platforms' community guidelines. Facebook's Community Standards were most comprehensive and covered all 66 types of abuse. YouTube's Community Guidelines came in second in terms of comprehensiveness, covering 56 out of 66 types of abuse. Discord's Community Guidelines, on the other hand, covered only 18 types, the least of the 11 platforms.

The analysis revealed some high-level patterns in the coverage of behavior types. All platforms had rules against adult non-consensual intimate imagery (commonly known as "revenge porn"), child exploitation imagery (commonly known as "child porn"), and minor sexualization. These infractions are severe and punishable by law— for example, child pornography is explicitly illegal in 94 countries (*Child Pornography*, 2015). Even within the U.S., 46 states in the U.S. already have established laws against revenge porn (Cyber Civil Rights Initiative, n.d.).

All 11 platforms also had rules against harassment and bullying, as well as inauthentic behavior, which generally refers to misrepresenting one's identity in order to mislead users or the platform. The widespread inclusion of harassment policies shows a heightened focus on the increasingly severe problem on social media; it is also a marked improvement from Pater et al.'s (2016) analysis of platform harassment policies, in which they noted that Twitter and Pinterest did not have explicit harassment policies in their community guidelines.

The inclusion of Inauthentic Behavior was the result of the increasingly common fake accounts that aim to spread false information or propaganda. Facebook, for example, regularly tracks and takes down multiple inauthentic accounts working in concert to mislead people and cause harm, a kind of abuse that Facebook names "coordinated inauthentic behavior" (Facebook, 2019).

On the other hand, some behavior types less commonly appeared in the community guidelines. For example, only Facebook had rules against human organ sale, and only two platforms (Facebook and Instagram) had rules against live animal sale. While also under the category of regulated goods, their coverage was significantly lower than non-medical drug sale and pharmaceutical drug sale (i.e., prescription drug sale), two arguably more common types of regulated goods on social media for which 9 platforms had rules. Also, only two platforms, Facebook and LinkedIn, specifically prohibit the celebration and promotion of one's own crime, but it is also possible that other platforms did not have rules in such granularity and conflated it with other high-frequency rules such as inciting violence, for which 10 platforms had rules.

Here I would like to note that these community guidelines are not static; instead, they evolve over time and go through frequent revisions. Just like how the harassment policies have changed since Pater et al.'s (2016) analysis, it is likely that the community guidelines have become more comprehensive since my analysis in November 2019, by covering either additional rules that I identified, or completely new rules not listed in the current content analysis.

While there is clear variability in the coverage of rules on different platforms, it is unclear why such variability exists, but we can speculate that platforms may have chosen to focus on and made rules regarding the types of misbehavior that is most rampant on their platform, or made explicit the rules that are most reflective of their values. My analysis shows that platforms indeed make different choices in terms of the acceptable and unacceptable behavior to make explicit. While no moderation system is perfect, these choices and trade-offs in rule making may contribute to the challenges and problems that platforms face.

Overall, my analysis in Phase 1 provides a comprehensive taxonomy of rules on social media platforms based on their published community guidelines. This taxonomy serves as the basis for me to investigate the severity of each type of abusive behavior, which I describe in Phase 2.

4.2 Phase 2: Perception Survey

The Phase 2 of this research aims to quantitatively analyze the qualitatively-produced taxonomy in Phase 1. Specifically, having identified 66 types of abusive behavior based on community guidelines of 11 major social media platforms in Phase 1, in Phase 2, I proceeded with developing the survey to solicit people's perceptions of the severity of them. The survey results will shed light on the study's ultimate goal of understanding global similarities and differences in the perceptions of abusive behavior.

Survey Development

Given that the goal of the study is to uncover global perspectives of abusive behavior online, I designed the survey with two primary goals: (1) generalizable across platforms, and (2) globally representative.

Survey Questions. To ensure the generalizability of survey results, in the survey I asked the participants to rate each type of abusive behavior identified in the cross-platform content analysis in Phase 1. Because participants may not adequately understand the particular names of each type of abuse, I asked participants to consider and rate the abuse in the context of a hypothetical scenario, where they needed to imagine someone violated a particular rule. For example, participants rated the severity of revenge porn by responding to the hypothetical scenario where they see someone posted photos of revenge porn. I also described the scenarios as if they happened on Facebook for two

reasons: First, my analysis in Phase 1 showed that Facebook's Community Standards was the only one that covered every type of abusive behavior identified. Second, concretely grounding the scenarios in a single platform could also mitigate the potential platform-related variability in the same abusive behavior (e.g., people may perceive harassment differently on Facebook than on Twitter).

While Likert scales as a type of ordinal scale are common for quantifying respondents' opinions in survey design, they suffer from well-known limitations such as the assumption that response options are equidistant and the tendency to cause anchor effects (Guilford, 1954). Using a bounded scale that typically has only five or seven options would also eliminate the nuances between 66 types of abuse, and therefore would limit my ability to model an accurate relationship. Therefore, I asked participants to directly report their perceived severity of abusive behavior through free-text numerical values. In other words, participants could freely input a number as their answer.

As the severity of abusive behavior can be an unfamiliar concept to regular users, in the survey I operationalized severity in two ways: punishment and urgency. The construct of punishment represents the negative consequences deserved for a certain type of abuse, which roots from the concept of proportionality in criminal justice literature: the more severe an abuse is, the graver the associated punishment should be (von Hirsch, 1992). The construct of urgency addresses platforms' problem of prioritization, which in practice guides the operation of many emergency response systems (Fiedrich et al., 2000): the more severe an abuse is, the sooner it should be taken care of.

To measure these two constructs, inspired by the willingness-to-pay measurement in economics research (R. C. Mitchell et al., 1989), I asked participants to respond with the amount of

money that (a) they would fine the person for conducting the abusive behavior and (b) the social media company should spend to remove the abusive behavior immediately over all other types of abusive behavior. In answering these two questions, participants could input any number as low as zero (which indicated that they believed the behavior was not abusive), and as high as they would like. To compare the results between the free-text numeric measurements and traditional Likert-scale measurements, I asked participants to rate the severity in a Likert scale question that measured how upset the participant would be upon seeing the abusive behavior.

As such, using the example of the behavior of sexualizing minors (termed "minor sexualization" in my analysis), below shows the corresponding scenario and the accompanying survey questions that a participant would see:

Imagine you saw:

A photo of **a minor in a sexual pose** on Facebook.

- How much money, if any, would you fine the person who posted this content? Please indicate your answer in {local currency}. You only need to enter a number. (Punishment measurement, free-text numeric)
- How much money, if any, do you think Facebook should spend to remove this content immediately over other types of content? Please indicate your answer in {local currency}. You only need to enter a number. (Urgency measurement, free-text numeric)
- How upsetting is this content to you, if at all? (Upsettingness measurement, Likert scale: {extremely, very, somewhat, a little, not at all} upsetting)

Recruitment. For the survey results to be globally representative, I strategically recruited survey participants from ten different geographical regions through Qualtrics panels, based on

published statistics of countries with the most Facebook users (Clement, 2019a). I recruited 2,128 participants in total, with approximately 200 participants per region (see Table 4-1 for the exact breakdown). This sample size was the result of budget constraints, having approximately equal representation per region, and having a sample size with a reasonable margin of error. For example, India, which had the largest Facebook user base, had 269 million Facebook users as of October 2019. Therefore a sample size of 200 could give me a satisfactory 7% margin of error with a 95% confidence level.

The participation criteria were that the participant was over 18 years old, and had used Facebook in the last 30 days. To prevent participants from lying in order to qualify for the survey, I asked participants to choose all the social media platforms that they used from a list of options, but they would only be qualified if they had chosen Facebook as an option, because the survey scenarios were described in the context of Facebook. Following recommendations in survey methodology literature (Cibelli, 2017), I also had participants answer a commitment question in the beginning that asked whether they agreed to thoughtfully provide their best answers, and would be screened out if they could not promise to do so. I also collected demographic data including age range, gender, and education level at the end of the survey to prevent early drop-off. The full survey instrument is shown in Appendix B.

Piloting and Translation. I piloted the survey with 36 people to test the validity of my approach. All pilot participants were able to understand the questions correctly and provide reasonable answers. However, according to their feedback, asking about all 66 types of abuse would make the survey overly lengthy, with a completion time of approximately 30 minutes. Therefore, to prevent fatigue and drop-off, in the final version of the survey, each participant saw a random 33, or

GEOGRAPHICAL REGION		LANGUAGE	# PARTICIPANTS	
Brazil		Portuguese	211	
Egypt		Arabic	207	
Europe	UK	English	72	215
	France	French	71	
	Germany	German	72	
India		Hindi	213	
Indonesia		Indonesian	217	
Latin America	Mexico	Mexican Spanish	44	217
	Argentina	Spanish	44	
	Colombia	Spanish	44	
	Peru	Spanish	43	
	Chile	Spanish	42	
The Philippines		Filipino	220	
Turkey		Turkish	204	
United States		English	215	
Vietnam		Vietnamese	209	

Table 4-1. Recruitment and translation details for each country in the study.

half of the 66 total types of abusive behavior. The randomization effectively made each type of abuse receive approximately 100 ratings per region, which still provided reasonable generalizability.

Finally, since many parts of the world do not speak English as the dominant language, I also had the survey professionally translated into the corresponding dominant language spoken in each geographical region. The {local currency} placeholder was also translated into the country's

dominant currency. For example, U.S. participants would see "U.S. Dollars (USD)," while Vietnamese participants would see "đồng Việt Nam (VND)"[1]. Table 4-1 shows the final list of regions where I deployed the survey, as well as the languages into which the survey was translated.

Data Cleaning

First, I had Qualtrics perform basic data quality checks to filter out obviously poor responses or non-responses. The checks included filtering out duplicate respondents, response time that was too long or too short, patterned responses such as all zeros ("straight-liners"), as well as device IP address lookups to ensure the respondent was physically in the country intended.

I then conducted my own data cleaning, which involved two steps:

- **Winsorization**: Because some participants entered spuriously high values for highly severe abusive behavior (e.g., a $\$10^{17}$ fine), I winsorized the free-numeric responses at 95% level (i.e. capping all values in the top 5% at the value at the top 5% point) to reduce the effect of extreme outlier values.
- **Normalization**: Because the absolute monetary values varied depending on each participants' own conception of money, as well as the currency being used, these raw values were not comparable across different people or across countries. Therefore, using min-max scaling, a common normalization technique that scales raw values within a certain range, I normalized the monetary values to a maximum of 550,000, which is the median of the maximum values that the participants entered. Specifically, I used the following normalization formula:

[1] Vietnamese for "Vietnamese Dongs"

$$x_{normalized} = \frac{(x - \min(x)) \times 550000}{\max(x) - \min(x)}$$

where x was each participant's response to the 33 types of abuse that they were presented, represented as a vector.

The normalization scaled all participants' response values to the same range of values, and therefore made them comparable with each other for my analysis. It also preserved the relative relationships between the numerical ratings for each participant.

4.3 How Do People in Different Regions Perceive the Severity of Abusive Behavior?

To start, I first explored how participants in each region viewed abusive behavior broadly, and how these regional perspectives compare with each other when juxtaposed. Because one of the goals of the study was to understand how different types of abusive behavior should be prioritized, I set out to examine how each region prioritized abusive behavior, which was reflected by its ranking. Given that each region also quantitatively rated the severity of each type of abuse, I also investigated how fast the severity values grew as the behavior got worse, as a rank order is inherently linear and may not reflect the true relationship.

To answer these questions, for each country, I first calculated the mean punishment and urgency values for each type of abusive behavior across participants, then calculated the overall severity value of each type of abuse as the mean of its punishment and urgency values. In other words, if s_k is the severity value for abusive behavior k, then

$$s_k = \frac{\overline{P_k} + \overline{U_k}}{2}$$

where $\overline{P_k}$ is the mean punishment values for k across participants, and $\overline{U_k}$ is the mean urgency values for k across participants. I then plotted the overall severity values of each type of abusive behavior against its rank order by region.

As shown in Figure 4-1, I observed exponential growths in the severity of abusive behavior consistently across regions.

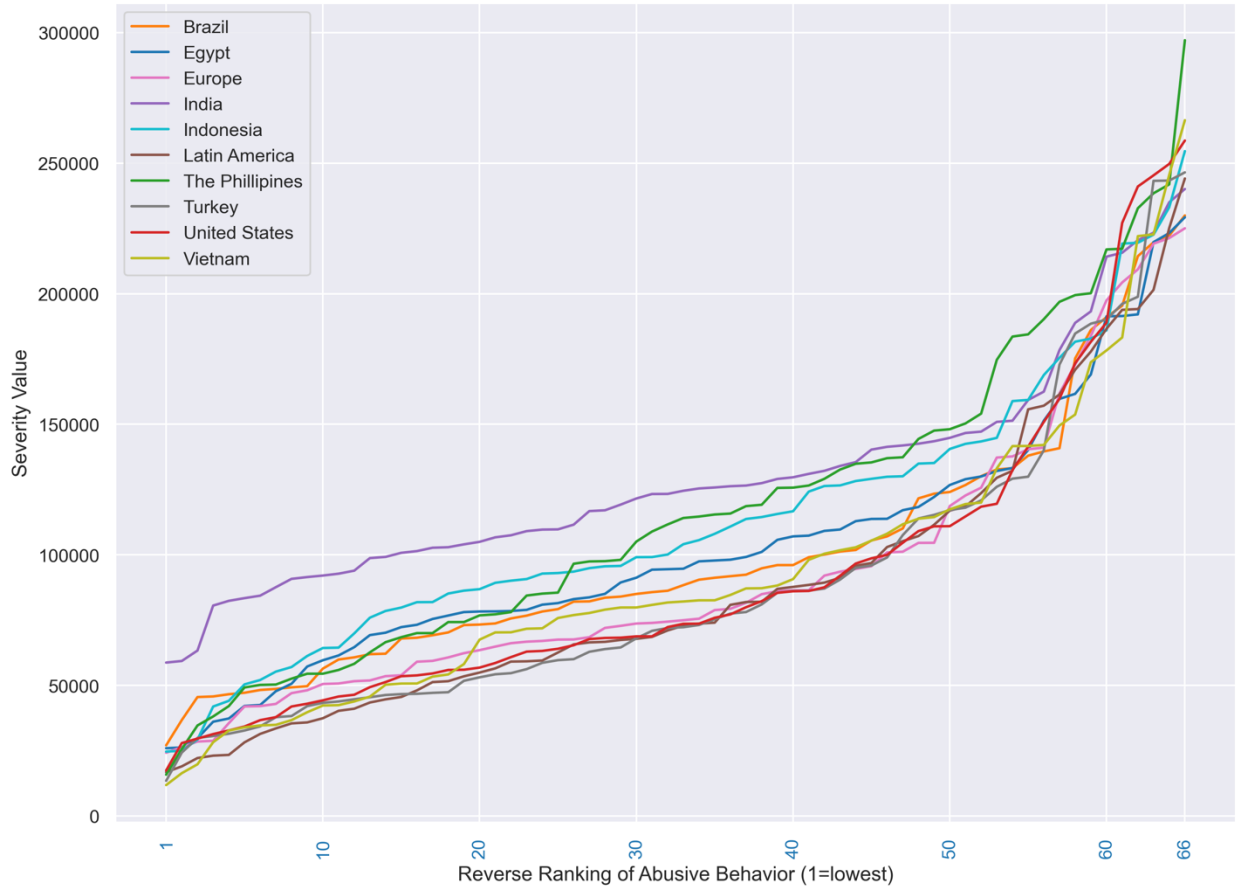

Figure 4-1. Plot of severity value vs. reverse severity rank order for each region under a free-text numeric measurement. Note that the same rank order may indicate different abusive behavior for different countries.

To confirm the exponential growth, I conducted an exponential regression on each region's data. Table 4-2 shows the regression results using the exponential function $y = Ae^{kx}$, where y is the severity value for each type of abusive behavior, and x is the reverse rank order of each type of abusive behavior (i.e., 1=lowest, 66=highest, in the same order as the x-axis in Figure 4-1). Here, k determines the growth rate of severity: The higher k is, the faster the severity value grows as the type of abuse becomes more severe. A represents the "starting point" of severity: The higher A is, the larger the severity value is for the least severe type of abuse.

REGION	A	k	R^2
Brazil (BR)	40878.84	.023	.944
Egypt (EG)	39838.25	.024	.909
Europe (EUR)	32120.74	.026	.944
Latin America (LATAM)	24354.06	.032	.965
India (IN)	72922.22	.015	.918
Indonesia (ID)	43873.91	.025	.894
The Philippines (PH)	37959.63	.028	.931
Turkey (TR)	26118.68	.030	.950
United States (US)	28545.27	.029	.949
Vietnam (VN)	27099.36	.031	.914

Table 4-2. Exponential regression results of each region's data. Here, p = .000 for all parameters.

As shown in Table 4-2, the regression analysis found that all parameters were statistically significant, thereby confirming the exponential growth across regions. In my sample, Latin America had the highest rate of growth but the lowest "starting point" of severity, while India had the lowest

growth rate but the highest "starting point." The consistent exponential growth means that, contrary to what a simple rank order would show, the "distances" between different types of abusive behavior were not equal—the "distance" between what was ranked in the first place and the second place was much larger than that between the 65th place and the 66th place. While a rank order would imply a steady linear severity growth as a result of assigning consecutive integers, the exponential growth revealed that as the rank grew higher, not only did the behavior become more severe, it also gained severity more quickly.

In order to ensure my free-text measurements' efficacy, I also tested the traditional Likert scale measurement as a comparison to the punishment and urgency free-response measurements, and I observed distinctly different relationships. Figure 4-2 shows a plot of the mean response of each type of abusive behavior vs. its reverse rank order by region; the only difference with Figure 4-1 is the y-axis represents the Likert scale measurement, rather than the free-response measurement. As Figure 4-2 indicates, while the growth rates of severity still seemed consistent across regions, they were clearly closer to linear growth instead of exponential growth. A possible reason for the distinctive difference in growth rates is the inherent upper limit of Likert scales—in this study, the most severe possible option participants could select was "extremely upsetting," thereby disallowing free growth. By limiting the higher end of severity, the Likert scale may also flatten the nuanced relationships between different types of abusive behavior, which the free-response measurements were able to reveal.

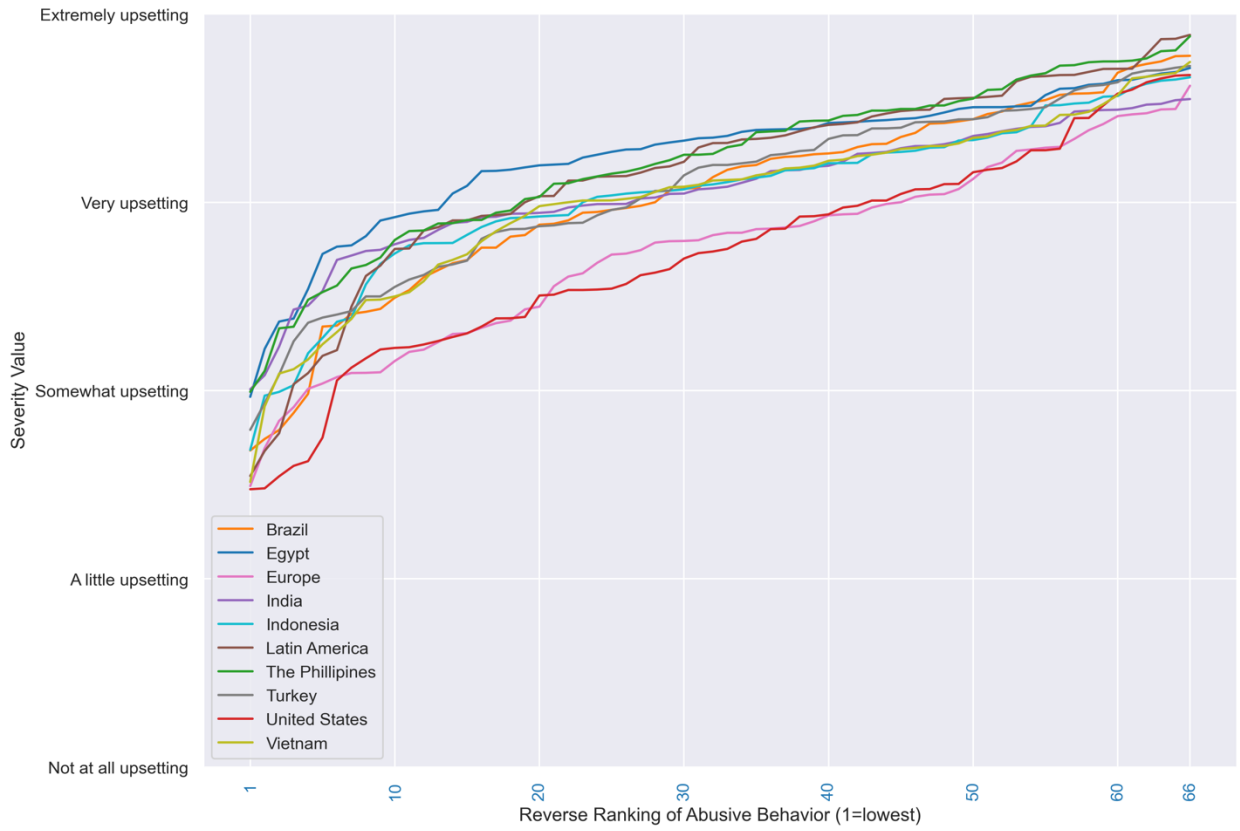

Figure 4-2. Plot of severity value vs. reverse severity rank order for each region under a Likert-scale measurement.

Overall, across the regions I examined, people's perceptions of abusive behavior's severity showed a distinctive pattern under a free-text numeric measurement: as the behavior became worse, the perceived severity grew exponentially. Furthermore, the commonly-used Likert scale measurement concealed the exponential growth by showing a linear growth instead, which would have underestimated the severity of many types of abusive behavior, especially those on the higher end of severity.

Despite the consistent exponential growth as the rank order of abusive behavior increases, it is important to note that the specific abusive behavior in any rank position varied from region to

region. In other words, participants in each region had their own ideas of how to rank different abusive behavior. I discuss these regional differences and similarities in the following sections.

4.4 What Are the Similarities and Differences Between Regions?

Overall, no two regions were the same. There was a lot more disagreement than agreement between countries, and there was always something that any two countries (strongly) disagreed on. However, I was able to observe clusters in which countries were more likely to agree with each other.

To identify clusters of countries that are similar, I first performed principal component analysis (PCA) on the rankings of abusive behavior by region. PCA is a common technique to reduce dimensionality in a dataset (Jolliffe, 2002), and was necessary in my analysis because there were much more dimensions (i.e., the number of abusive behavior types) than data points (i.e., the number of regions), which would make for more expensive computation and less meaningful clusters (Kriegel et al., 2009). I experimented with the number of components $n \in [2,9]$ and found that when I projected the original 66-dimensional data onto $n = 7$ components, the result could explain

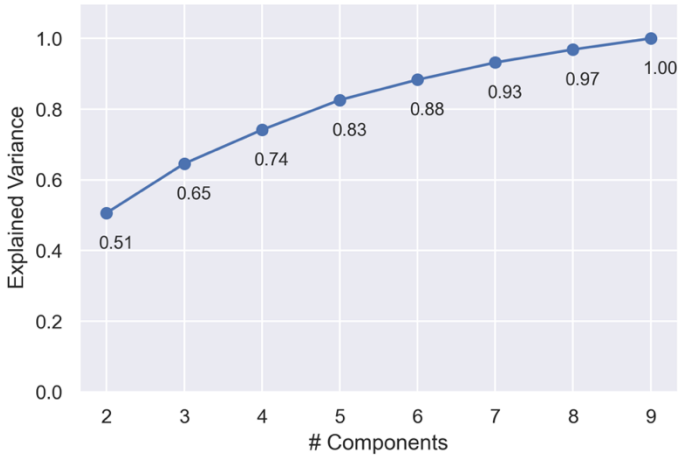

Figure 4-3. Plot of explained variance vs. number of principal components in PCA.

93% of the original variance (shown in Figure 4-3). Therefore, I chose $n = 7$ to keep the number of components to a minimum, while not sacrificing too much information from the original data.

After reducing the number of dimensions from 66 to 7 using PCA, I then used the *K*-means algorithm (Hartigan & Wong, 1979) to find clusters. I determined the number of clusters *K* by examining the average silhouette coefficient *s* for each possible *K*, which represents how well samples are clustered with samples that are similar to themselves (Rousseeuw, 1987). The higher the silhouette coefficient *s*, the better the clustering is.

I tested the *K*-means algorithm by experimenting with $K \in [2,9]$, and found that the best clustering appeared at $K = 3$ ($s = .18$) and $K = 4$ ($s = .17$), before the silhouette coefficient started rapidly decreasing (shown in Figure 4-4). The difference between $K = 3$ and $K = 4$ was whether or not Turkey belongs to the same cluster as Vietnam and Indonesia. While the model achieved the highest silhouette score at $K = 3$, upon qualitatively examining the clusters, I found that the ranking of Turkey showed a unique pattern that was different from Vietnam and Indonesia (explained below). Therefore, I proceeded with $K = 4$ clusters, even though it did not technically have the highest silhouette score.

Figure 4-4. Plot of average silhouette score vs. number of clusters in K-means clustering.

To confirm the robustness of my clustering, I also conducted a pairwise ranking correlation analysis; Table 4-3 shows the correlation results. The analysis also showed the same clustering, with within-cluster correlations consistently higher than 0.8. In other words, the rankings of regions within each cluster are highly correlated with each other, which corroborates with the *K*-means clustering result.

	BR	EUR	LATAM	US	IN	PH	EG	VN	ID	TR
BR	1.0000	.8835	.8889	.8764	.6956	.8138	.6641	.7310	.6790	.7660
EUR	.8835	1.0000	.8352	.8725	.6954	.7784	.6713	.6822	.6726	.7697
LATAM	.8889	.8352	1.0000	.8172	.7471	.8278	.6751	.7827	.7076	.7049
US	.8764	.8725	.8172	1.0000	.7491	.8421	.7407	.6503	.7017	.7623
IN	.6956	.6954	.7471	.7491	1.0000	.8910	.8215	.7606	.7652	.7609
PH	.8138	.7784	.8278	.8421	.8910	1.0000	.8094	.7586	.8067	.7943
EG	.6641	.6713	.6751	.7407	.8215	.8094	1.0000	.6336	.6992	.6729
VN	.7310	.6822	.7827	.6503	.7606	.7586	.6336	1.0000	.8080	.7557
ID	.6790	.6726	.7076	.7017	.7652	.8067	.6992	.8080	1.0000	.7697
TR	.7660	.7697	.7049	.7623	.7609	.7943	.6729	.7557	.7697	1.0000

Table 4-3. Pairwise ranking correlation results for all regions.

Each cluster had its own set of behavior that it perceived as more or less severe compared to other clusters, based on how they ranked these types of behavior. Table 4-4 lists the behavior perceived as more or less severe by each cluster.

Overall, I found that each region cluster had a unique set of abusive behavior that they collectively perceived as more severe or less severe, and these sets rarely overlap between clusters. There are also types of behavior that are ranked highly by one cluster, but low in another (e.g.,

marijuana sale has a high ranking in Cluster 4 but low ranking in Cluster 1). Furthermore, while Turkey would have been in the same cluster as Vietnam and Indonesia had I strictly followed the highest silhouette score, it had no overlap with them in either the highly ranked or the lowly ranked behavior. The non-overlap and the opposite ranking of certain behavior show the necessity of treating different regions differently in creating rules and policies, rather than taking a one-size-fits-all approach that is likely to deprioritize abusive behavior certain regions perceive as highly severe.

CLUSTER	BEHAVIOR PERCEIVED AS MORE SEVERE	BEHAVIOR PERCEIVED AS LESS SEVERE
Cluster 1 Brazil (BR), Europe (EUR), Latin America (LATAM), United States (US)	Animal Abuse Mutilated Humans Child Nudity	Marijuana Sale Adult Nudity Theft Prostitution Sexual Activity
Cluster 2 India (IN), The Philippines (PH), Egypt (EG)	Sadism/Glorifying Violence Sexually Explicit Language Adult Nudity Sexual solicitation Human Organ Sale Coordinating harm Adult Non-Consensual Intimate Imagery	Child Exploitation Imagery Child Abuse Animal Abuse Mutilated Humans
Cluster 3 Vietnam (VN), Indonesia (ID)	Hate Speech: Dehumanization Self Injury Promotion	Child Exploitation Imagery Inappropriate Interactions with Children Child Nudity Minor Sexualization Creep Shots
Cluster 4 Turkey (TR)	Harassment Eating Disorder Promotion Commercial spam Marijuana sale	Criminal Group Coordination Criminal Group Propaganda Non-consensual Sexual Touching Digital Nudity Suicide Depiction

Table 4-4. Region clusters and their ranking characteristics.

4.5 On What Abusive Behavior Do Regions Agree and Disagree?

While I have examined the patterns and differences on the level of geographical regions in terms of how they perceive abusive behavior, it is also valuable to conduct a similar investigation on the level of behavior. Insights into the consensus and disagreements within individual behavior types will reveal which types of behavior may deserve more nuanced treatment than others.

To examine agreements and disagreements, I measured the level of agreement using abusive behavior's max ranking differences. Here, I define the max ranking difference ($\Delta rank$) within a type of abusive behavior as its lowest ranking (i.e., highest in number) across regions less the highest ranking (i.e., lowest in number) of that abuse. Note that the rank order here has 1 as the highest rank, as opposed to the x-axes in Figure 4-1 and Figure 4-2.

While there are other ways to measure disagreement such as standard deviation (SD) and interquartile range (IQR), I did not use them for specific reasons. I did not use SD because it tends to increase with the ranking values, which is likely to cause lower-ranked abusive behavior to generally have larger SD. I also did not use IQR because it aims to exclude "outliers" in the data (i.e., those that are in the first and the last quartile), but it would be undesirable to exclude "outliers" here—excluding a whole region simply because it ranked differently would defeat the purpose of considering diverse regional perspectives.

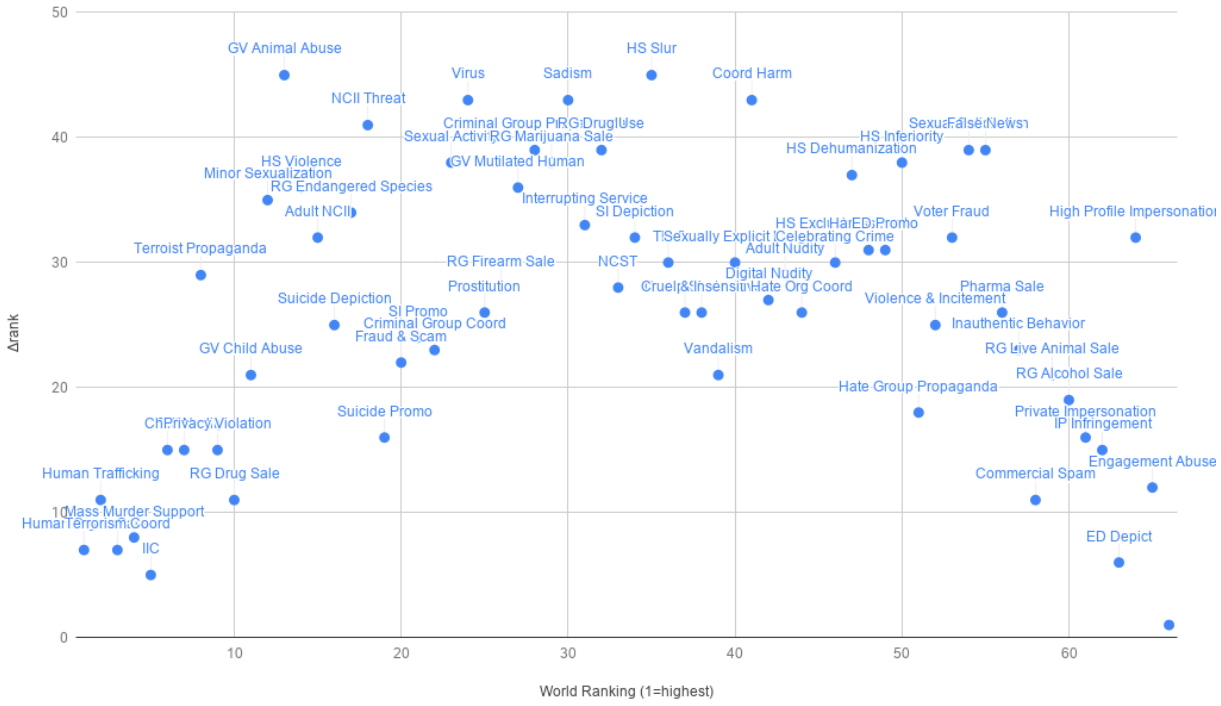

Figure 4-5. Plot of each type of abusive behavior's max ranking difference (*⊿rank*) across regions vs. its overall world ranking.

The analysis showed that the largest Δ*rank* existed in minor sexualization and self injury depiction, both with a Δ*rank* = 45. The smallest Δ*rank* existed in engagement abuse (which often refers to clickbaits), with a Δ*rank* = 1. Figure 4-5 shows all 66 types of abusive behavior's Δ*rank* plotted against their overall world severity ranking (1=highest). Note that the overall world ranking here is a region-agnostic one, rather than the aggregate of by-region rankings.

The reverse-U shape in Figure 4-5 shows that regions agreed more on the most severe and the least severe behavior, but disagreed more toward the middle. To take a closer look at larger disagreements, I counted 19 types of abusive behavior whose Δ*rank* are at least 33, or half of the total number of rank positions, as shown in Table 4-5.

ABUSIVE BEHAVIOR	*Δrank*
Minor Sexualization	45
Self Injury Depiction	45
Adult Sexual Activity	43
Regulated Goods: Marijuana Sale	43
Sexually Explicit Language	43
Regulated Goods: Endangered Species Sale	41
Graphic Violence: Mutilated Humans	39
Interrupting Platform Services	39
Voter Fraud	39
Sexual Solicitation	39
Criminal Group Coordination	38
Criminal Group Propaganda	38
Eating Disorder Promotion	38
Celebrating Crime	37
Graphic Violence: Animal Abuse	36
Regulated Goods: Firearm Sale	36
Graphic Violence: Child Abuse	35
Suicide Depict	34
Sadism	33

Table 4-5. Types of abusive behavior that had max ranking differences (*Δrank*) of at least 33, or half of the total number of rank positions.

While there is not a clear pattern in the kinds of abusive behavior listed in Table 4-5, the fact that 19 types of abuse, or nearly 30% of all identified types of abuse, received large disagreements across regions shows the necessity of customizing content moderation by geographical regions—an umbrella approach is likely to mistreat a nontrivial amount of abusive behavior by dismissing regional differences, and the behavior on which people heavily disagreed deserves greater attention, more in-depth research, and more nuanced treatment.

4.6 What Are Some Regionally Sensitive Topics?

The region-level and the behavior-level analysis in the previous two sections revealed that disagreement in the perception of abusive behavior widely existed. While there were clear clusters in regional perceptions, I found less patterns when examining individual types of abusive behavior. Therefore, in an effort to further seek out patterns in the kinds of behavior that regions agreed or disagreed on, my colleague and I qualitatively categorized individual types of abusive behavior into higher-level topic categories.

We first independently categorized all 66 types of abusive behavior on our own, with an eye toward achieving a reasonably small number of categories so we can easily identify patterns. We also tried to follow how regular people rather than moderation experts would interpret abusive behavior. For example, while Facebook's Community Standards categorized human organ sale as part of regulated goods, an average person may associate it with mass scale killing, and in our categorization we followed the latter rationale. We then came together to discuss and adjudicate differences, while iterating on the categories. While I acknowledge that our primarily U.S.-centered perspective is likely to have had an influence on our categorization, I believe it is still a promising first step to a preliminary understanding of the similarities and differences in a global context. We eventually agreed upon the following topic categories:

- Mass Scale Harm (e.g., terrorism, human organ sale)
- Vulnerable Groups (e.g., child exploitation imagery, child abuse)
- Violence (e.g., graphic violence: mutilated humans, sadism)
- Platform Abuse & Spam (e.g., interrupting platform services, engagement abuse)

- Sexual Violence / Sexual Content (e.g., sexual activity, sexual explicit language)
- Regulated Goods (e.g., marijuana sale, drug sale)
- Self-harm (e.g., suicide depiction, eating disorder promotion)
- Financial harm (e.g., fraud & scam, privacy violation)
- Other Directed Harm (e.g., theft, vandalism)

Using these categories, I then generated a color map (shown in Table 4-6) by mapping the individual types of abuse to the above categories with color-codings. The color map resonated the findings in Figure 4-5: more consistency on the upper and the lower ends, but less in the middle. Below, I describe some high-level patterns in the most and the least severe abusive behavior rated by participants.

Abusive Behavior That Were Most Severe

Overall, abusive behavior involving mass scale harm and vulnerable groups had a large share of what participants perceived as most severe. For all regions, the top 4 types of abusive behavior were consistently about mass scale harm and vulnerable groups, which all had to do with: mass murder, child exploitation, human organ sale, human trafficking, and terrorism. Furthermore, the top 10 types of abuse rated by participants from countries in Cluster 1 (Brazil, Europe, Latin America, United States) were almost completely mass scale harm and vulnerable groups; the only exception was Europe, whose 10th ranked behavior was in the violence category (Graphic Violence: Mutilated Humans).

In addition to mass scale harmand vulnerable groups, regulated goods was also a common category for regions outside Cluster 1. For participants in Indonesia, Egypt, and Turkey, drug sales were the only abuse related to regulated goods that made into the top 10. Vietnam participants,

however, rated three types of regulated goods—drug sale, endangered species sale, and marijuana sale—among its top 10.

Additionally, financial harm only appeared in Cluster 2 (The Philippines and India) among its top 10, and sexual content appeared in the top 10 of only Cluster 2 and 3 (Egypt, India, Indonesia). Violence only appeared in the top 10 of two regions, Indonesia and Europe. Self-harm, platform abuse, and other directed harm were not among the top 10 in any region.

Abusive Behavior That Were Least Severe

Compared to the most severe types of abuse, the least severe types showed less patterns. Platform abuse appeared frequently in the lowest-ranked abusive behavior in terms of severity. Specifically, engagement abuse, which refers to posting clickbait-like content, was consistently in the last place except Vietnam and Egypt, where it was second to last. Intellectual property infringement and commercial spam were also consistently in the bottom 10.

Regulated goods was also a common category in the bottom 10, but only included live animal sale, alcohol sale, and prescription drug sale. Surprisingly, for 7 out of the 10 regions, eating disorder depiction was in the bottom 10, despite being a dangerous mental illness. The only exceptions were Egypt, Turkey, and Europe.

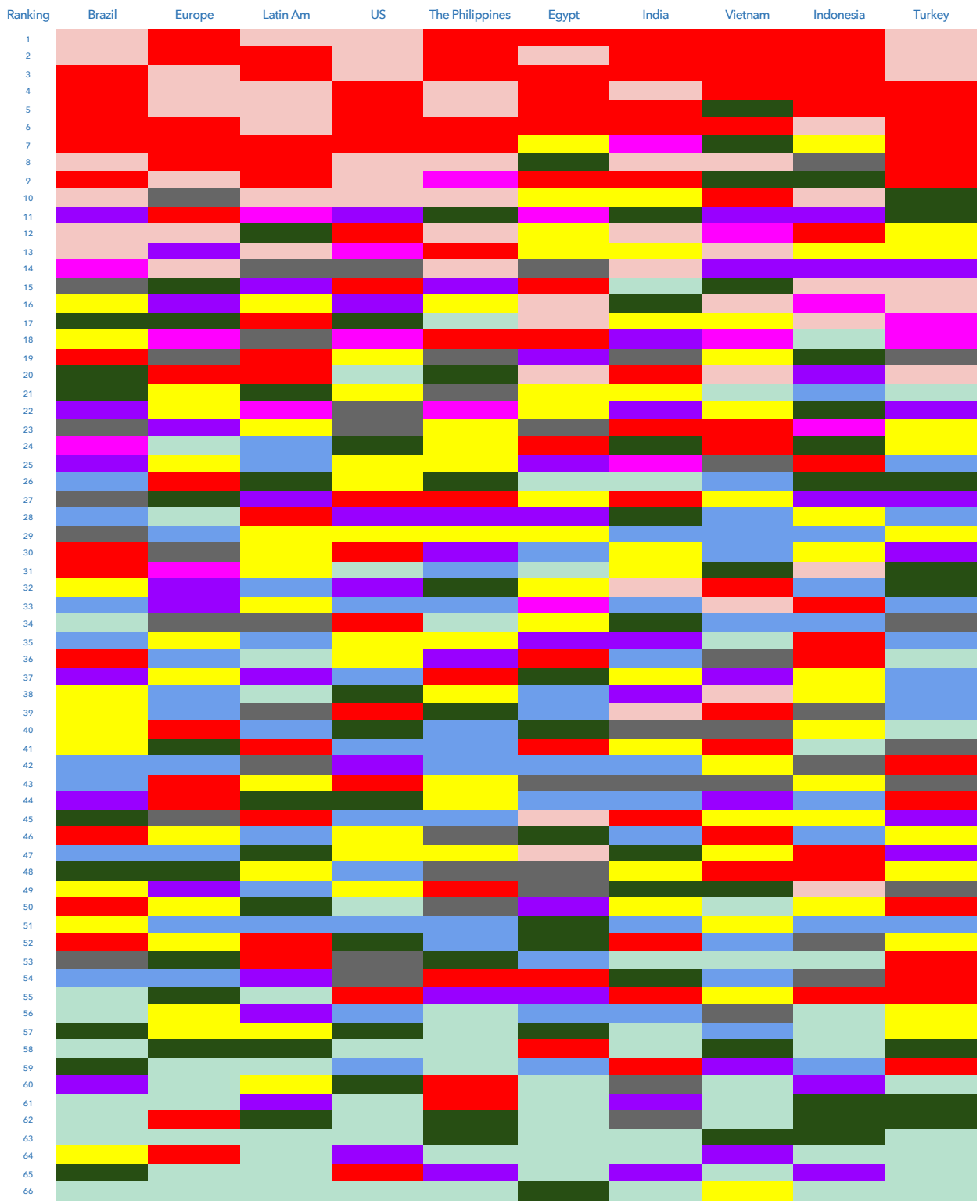

Table 4-6. Color map of abusive behavior topic categories, ranked from high to low by region.

4.7 Discussion

Overall, my findings show significant differences in perceptions of abusive behavior across different regions in the world. The general variability serves as evidence against social media platforms' current approach of using a single set of rules to regulate global users, which falsely implies that people view abusive behavior consistently across the world. The global disagreement widely existed in different facets of my analysis—in different regions' rankings of abusive behavior, in individual types of abusive behavior, and also in the higher level topics that categorizes these individual types.

Despite the general variance, there were some patterns in how global users perceived abusive behavior. First, regardless of how each region ranks abusive behavior, I found that the perceived severity followed the same type of growth as the abusive behavior became increasingly worse—specifically, an exponential growth. Second, while there are no unified ways to describe how all regions ranked abusive behavior, there were clusters within which the rankings were similar, which suggests the possibility of moderation by country or region groups. Finally, in my analysis of the agreement and disagreement on the level of individual types abusive behavior as well as high-level topics, the convergence toward the extreme ends suggests that people had more consensus on the highly severe types of abuse as well as the highly non-severe types of abuse, but had larger differences of opinions toward the others in the middle.

Taken together, my findings have several implications for platform content moderation. First, the consistently exponential growth across all regions provides a principle for moderation prioritization in localized contexts—while what kinds of abuse come first may vary from place to place, the underlying relationship between the types of abusive behavior may remain the same. The

distinctive exponential growth also shows that results from a simple ranking may be misleading because it implies a linear growth, and highly severe content may not be receiving the attention that it deserves.

Furthermore, my findings indicate that policymakers of social media platforms, who likely developed community guidelines from a predominantly Western (and particularly U.S.) point of view (Gillespie, 2018a), may have deprioritized abusive content that are perceived as highly severe in non-U.S. regions. However, I am not claiming that platform policy making should be in a "U.S. vs. elsewhere" fashion, because my evidence shows that on a global scale, perceptions of abusive behavior are complex, and highly varying even in regions outside the U.S.

The risk of careless categorization is even more pronounced given that the moderation resources and capacity are inevitably limited. For example, if a platform performs the same clustering analysis in this study and decides to place Turkey in the same cluster as Vietnam and Indonesia, it is likely that the high-severity types of abuse of these three countries will be fighting over the same pool of resources with some being deprioritized. Instead, a better approach might be for each cluster to have their own pool of resources taking care of their own priorities.

Therefore, I argue that a promising direction for content moderation is to customize and localize by regions. At the same time, I also acknowledge that differences may still exist within geographical regions, and more research is needed to uncover the specific nuances in any particular region.

My analysis of regionally sensitive topics revealed a similar pattern of consistent opinions about the most severe (mass scale harm and vulnerable groups) and the least severe (platform abuse and certain kinds of regulated goods) behavior, but diverse opinions about the others. While

platforms may want to highly prioritize behavior involving mass scale harm and vulnerable groups (as they may already be doing), it would be perilous to treat the least severe ones as "not important," because regular users may not be aware of the non-obvious harms that they could cause (e.g., sale of prescription drugs or intellectual property infringement). Instead, platforms may take the opportunity to educate users about the harms of these types of behavior perceived as less severe. Furthermore, the differences in the variance of the perceptions show a higher need of research understanding of the more diversely perceived behavior, as these types of behavior are likely to be more complex and simplistic moderation strategies for them may unintentionally privilege certain groups of people at the cost of others, a point echoed by Schoenebeck et al. (2020) in their study of different people's perceptions of justice models.

The consistencies and variances across the world in the perceptions of abuse reveal the complexity of content moderation, and the necessity of a multi-stakeholder perspective that this dissertation argues for as a whole. In Chapter 6, I will discuss the design and ethical implications of the research in this chapter in more detail, but in the next chapter, I will take a step back and holistically examine moderation literature, and uncover the tensions and trade-offs hidden in them.

5
making different choices in content moderation

The previous two chapters have showed how a multi-stakeholder perspective can reveal critical problems and perspectives that would remain unnoticed if we constrained our focus within a single stakeholder. While focusing on one group of stakeholders is reasonable and sometimes favorable for individual studies, it inevitably makes other stakeholders and contexts invisible. However, there are many examples where the same content moderation practices work differently for different people in different contexts: For example, automated moderation can work at the consistency and speed that large-scale moderation requires, but lacks nuanced understandings needed by individual cases that often fall into the gray area of policies and rules (Jhaver, Birman, et al., 2019). Moderators in text-based online spaces rely heavily on removing and editing content, but the same methods completely break down in communities where voice chat or virtual reality is the

dominant mode of interaction (Blackwell et al., 2019; J. A. Jiang et al., 2019). These two examples, and many more like them, demonstrate the complexity and difficulty of content moderation in practice, when there are always multiple stakeholders, multiple needs, and multiple contexts. An examination of content moderation research in these different dimensions will offer new insights into past moderation practices and challenges, and encourage researchers, designers, moderators, and regular internet users to reflect on content moderation by considering factors that they may not have considered before.

In this chapter, I present a trade-off-centered framework of content moderation, developed through a systematic literature review of 83 papers that document empirical studies. My framework is characterized by four major, interrelated trade-offs at increasing levels of abstraction: Trade-offs in moderation actions, styles, philosophies, and values. Every decision in these four categories has potential positive and negative outcomes. I first provide a detailed description of of each of the four layers in my framework, then I show how researchers, designers, and moderators can use my framework of trade-offs in their own work.

5.1 Method: Systematic Literature Review

To understand patterns and trends in existing literature about content moderation, I conducted a systematic literature review, following best practices established in different fields (Liberati et al., 2009), as well as rigorous review studies in the HCI and CSCW literature (Chancellor et al., 2019; DiSalvo et al., 2010; Seering et al., 2018). This section will describe my search strategy to identify candidate papers, inclusion criteria to filter the candidate papers into a corpus for analysis, and analysis techniques.

Search Strategy

Prior work in content moderation shows that there is not one field that completely covers all content moderation research. A published reading list of content moderation (Gillespie, 2019) shows this line of research primarily happens in social computing, human-computer interaction, computational social science, and communication, spanning a wide range of ACM and AAAI conferences (e.g., ACM CHI, ACM CSCW, AAAI ICWSM) and journals (e.g., Social Media + Society, New Media & Society, International Journal of Communication).

Therefore, to ensure a robust coverage across venues, I used a combination of search databases. Following methods used by Chancellor et al. (2019) who did a similarly interdisciplinary meta-review, I used the ACM Digital Library to search ACM journals and conferences, the AAAI Digital Library (implemented with Google custom search) to search AAAI publications, and Web of Science for other journal publications.

Using a keyword search within the above databases, I identified an initial set of candidate papers published between 1998 and the day I conducted the search, following one of the earliest documented misbehavior in cyberspace (Dibbell, 1998) that predated the social media era. Based on keywords used in all published papers about content moderation in CSCW 2018 and 2019, two venues with a relatively high amount of empirical content moderation work, as well as keywords that they have used to describe content moderation, supplemented with my domain knowledge, the final list of keywords included:

content moderation, platform moderation, community moderation, platform governance, community governance, internet governance.

In order to check the validity of these keywords, I manually went through every paper (regardless of whether it related to content moderation) published in one conference (ICWSM 2019) and one journal (Social Media + Society papers published in 2019)—which constituted a total of 158 published papers—and created a subset of papers about moderation. I then performed a keyword search of that conference and journal, and ensured that the keyword search did not result in any false negatives. False positives were retained, since they could be filtered out by my inclusion criteria (described below). My search strategy finally yielded 1,074 papers in total (309 from the ACM Digital Library, 35 from the AAAI Library, and 730 from Web of Science).

Inclusion Criteria

Each paper identified with my keyword search needs to meet the following criteria to be included in the corpus:

- **Archival & peer-reviewed:** A paper needs to be archival and peer-reviewed for inclusion, because these papers have been scrutinized by experts to ensure their validity and rigorousness and thus meet the publication criteria of the chosen venues. I did not include non-archival papers such as late-breaking work or workshop papers because they often include work that is incomplete and ongoing.
- **Empirical study:** The paper needs to describe at least one empirical study to be included for analysis. An "empirical study" here means a study that collects data from people. This definition means:
    - Since I focused on real-world moderation practices and challenges validated by real moderators, I did not include essays or papers that are purely

theoretical analysis. However, studies that use empirical evidence to validate social science theories would meet this criterion.

    - Papers that describe systems would only qualify if they also describe user studies, which includes formative studies before building the system, and evaluative studies of how people use the system.
    - I also did not include papers that only summarize or evaluate existing studies, because the qualifying studies that these papers build on would already be included in the keyword search.

- Moderation practices, challenges, impacts, or recommendations: For this study I only focus on these four facets of moderation. Therefore, for a paper to be included, it needs to document at least one of the following:
    - Existing moderation practices or approaches;
    - Existing moderation challenges or problems;
    - Impacts and consequences of existing moderation practices;
    - Recommendations, implications, or future directions for designing or implementing content moderation.

Since these details may not be included in the paper titles or keywords, I also read the abstract of each paper.

After manually filtering and deduplicating using the inclusion criteria above, I retained 71 papers (33 from the ACM Digital Library, 7 from the AAAI Library, and 31 from Web of Science). I further added 12 papers from reading the bibliographies in these papers, resulting in a total of 83

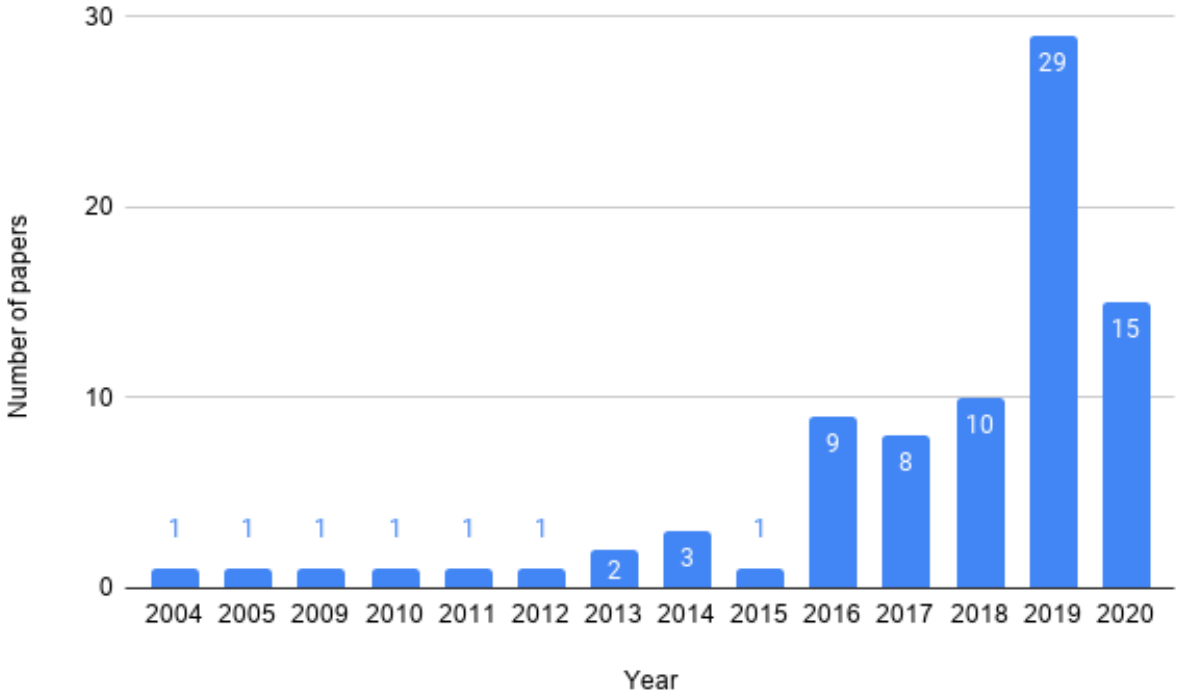

Figure 5-1. Number of papers in the dataset by year.

papers in my corpus listed in Table D-1 in Appendix D. Figure 5-1 shows the number of papers by year in my corpus.

Analysis Techniques

My research assistant and I conducted a thematic analysis of the papers in our corpus. We first engaged in one round of open coding by closely reading a sample of the corpus. Specifically, we each sampled one paper per year for every year with any publication, with an eye toward a breadth of research paradigms (e.g., qualitative and/or quantitative) and topics (e.g., volunteer moderation and/or commercial moderation), and open coded these papers. During this round of open coding, we regularly came together to discuss emergent code groups such as "moderation transparency" and "automated moderation bots." Then, we open coded the rest of the corpus with an eye toward the

preliminary code groups identified in the sample. Two more rounds of iterative coding led us to identifying higher-level categories such as "moderation actions" and "rules and norms." I used these categories to produce a set of descriptive theme memos that described each category grounded in the quotes from the papers. We then discussed the theme memo and developed the relationships between the categories, which resulted in the trade-off-centered themes that constitute the framework I discuss below.

5.2 A Trade-off-Centered Framework of Content Moderation

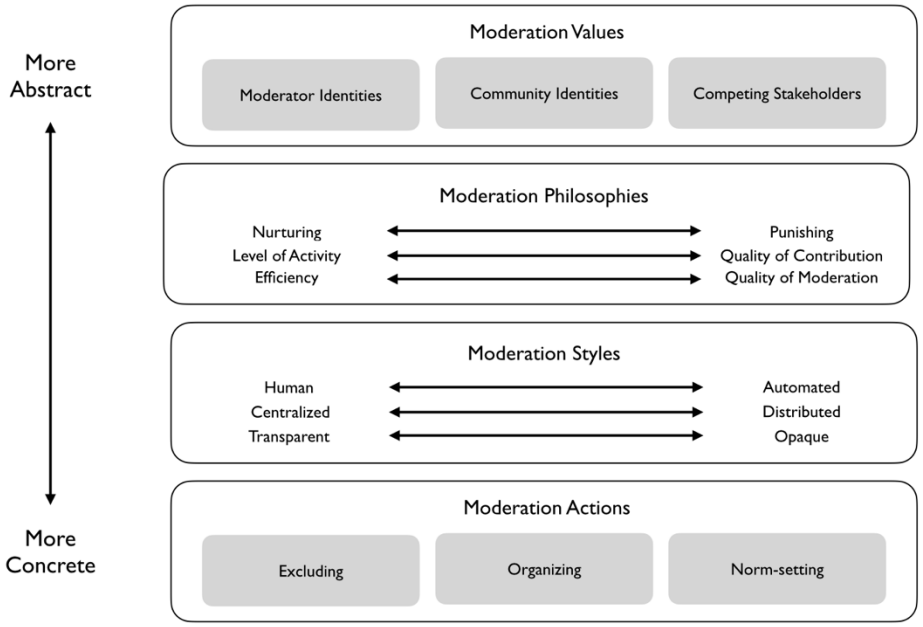

Figure 5-2. Diagram of my trade-off-centered framework of content moderation. The level of abstraction increases from moderation actions to moderation values. Note that elements and arrows within a single layer do not vary in the levels of abstraction.

My trade-off-centered framework of content moderation consists of four interrelated layers of trade-offs in increasing levels of abstraction: Moderation actions, moderation styles, moderation philosophies, and moderation values. Figure 5-2 shows a diagram that visualizes this framework.

Trade-offs in the more abstract layers impact those in the more concrete layers. I envision my framework to be an analytical tool that helps people examine and make sense of content moderation practices, rather than a mental model that prescribes moderators' thought processes.

The moderation actions layer represents multiple concrete moderation techniques that moderators can use to manage their communities, such as issuing warnings, removing content, and banning people. Following Grimmelmann's (2015) categorization of moderation "verbs," I describe these actions into three categories: excluding, organizing, and norm-setting. These actions had varying levels of harshness, and reveal trade-offs between stifling the community and exposing the community to harm, as well as in forgoing the opportunity to educate community members about acceptable behavior by immediately removing violating behavior.

The moderation styles layer goes up one level of abstraction by addressing how moderators can carry the actions in the moderation actions layer, similar to the "adverbs" in Grimmelmann's moderation framework. The styles layer consists of three specific trade-offs representing competing choices in the ways in which any of the moderation actions could be taken: human vs. automated, centralized vs. distributed, and transparent vs. opaque.

The moderation philosophies layer is one level more abstract than the moderation styles layer, and describes the philosophies that guide tendencies toward specific choices in moderation styles and moderation actions. The moderation philosophies layer consists of three trade-offs representing competing needs in content moderation: nurturing vs. punishing, level of activity vs. quality of contribution, and efficiency vs. quality of moderation.

The moderation values layer is the topmost layer, representing the competing values that impact decisions in the trade-offs in moderation philosophies, styles, and actions. I broadly classify

the values into three categories: Moderator identities, community identities, and competing stakeholders.

My analysis, which I describe later in the following sections, revealed increasing levels of abstraction in my framework: While prior literature often describes moderation actions and styles as concrete findings (to the extent that there are existing categorizations such as Grimmelmann's), moderation philosophies and values are more evasive, which are often discussed only as speculations, if being discussed at all. For the same reason, while I could easily find detailed investigation of the trade-offs in moderation actions and styles in individual papers, my discussion of the trade-offs in moderation philosophies and values required a synthesis of multiple papers.

How prior literature discussed the trade-offs also influenced my organization within the layers. The trade-offs in moderation styles and philosophies existed in clear, opposing binaries in my corpus, so I used arrows to represent them. However, trade-offs in moderation actions and values involved multiple possible options, and as a result, here I directly listed the categories of options instead. While I vertically order the layers to represent different levels of abstraction, it does not apply to the trade-offs within the moderation styles and philosophies layers; these trade-offs are equal and do not vary in the levels of abstraction.

It is important to note that none of the options in any of the four layers are mutually exclusive. Real content moderation practices are almost always a mixture of different options, with different actions, styles, philosophies, and values existing at the same time. Therefore, the arrows in moderation styles and philosophies are "slider scales" where the decision could fall anywhere in the middle, instead of at one extreme or the other. Similarly, choices in moderation actions and values

are also not mutually exclusive. My notion of a trade-off is not a one-vs-all choice, but a balance to achieve among many legitimate alternatives.

In the remainder of this chapter, I explain each of the four layers of trade-offs in detail, and close by discussing how different people can use my framework in their own work.

5.3 Trade-offs in Moderation Actions

The first trade-off that I identified was around the moderation actions against rule-breaking behaviors, similar to the techniques, or "verbs," in Grimmelmann's (2015) framework of content moderation. I found that moderators took different actions (for example, removing content or issuing warnings) to enforce content moderation. These actions had various levels of harshness, associated with different, sometimes competing outcomes and consequences.

Grimmelmann categorized techniques into four broad categories: excluding, organizing, norm-setting, and pricing. In my corpus, I also found more granular moderation actions that correspond to the first three categories. I did not find any direct evidence of pricing, likely due to social media platforms' overall pursuit of a high level of user engagement and lack of incentive to inhibit participation.

One of the common actions was exclusion, which means to deprive certain people of access to an online community, and often takes the form of banning and the less harsh version of it, timeouts (i.e., ban from participation for a certain period of time). 52 out of 83 papers in my corpus mentioned some type of exclusion. Sometimes whole communities may be excluded by platforms, such as the ban of several subreddits in 2015 due to their violation of Reddit's anti-harassment policy (Chandrasekharan et al., 2017). In communities with voice chatting functionalities,

moderators also practiced muting, which excludes people from participating in voice chats but not necessarily text chats (J. A. Jiang et al., 2019). The widespread use of exclusion was captured by Seering et al. (Seering et al., 2019), nearly all of whose moderators participants used exclusion in their work.

Organizing, appearing in 64 papers, was the most common type of action that focuses on content rather than people. It "shapes the flow of content from authors to readers" (Grimmelmann, 2015), which, in my corpus, includes removing and annotating content. While removal often intends to solely get rid of content that violates the community rules, annotating can serve a multitude of purposes. For example, post annotations in Reddit, called "flairs," are used as labels that categorize posts, whereas annotations in Wikipedia such as the Neutral Point of View (NPOV) tag are meant to notify readers that an article may be violating certain Wikipedia guidelines. In the case where the organized content is violating, prior research indicates differences between removing and annotating in the efficacy of helping community members adhere to norms. In a study of r/ChangeMyView, Srinivasan et al. (2019) showed the causal effect that post removals indeed improve norm adherence. In contrast, Pavalanathan et al. (2018) found that NPOV tags in Wikipedia did not help the editors to adopt the desired writing style, but did improve the overall quality of tagged articles, likely because of the contribution of other editors who edited upon the NPOV tags.

In addition to direct sanctions taken on people or content, moderators also widely used warnings (mentioned by 25 papers), which are less harsh, and fall into the premise of norm-setting by denouncing bad behavior (Grimmelmann, 2015). Moderators issued warnings to tell rule violators that they did something wrong, and sometimes also did so publicly to educate the

community more broadly. Seering et al. (2019) also noted that warnings ranged from light to strong, the latter often accompanied by temporary sanctions mentioned above. Skousen et al. (2020) in their study of an online health community also documented "indirect policing" practices similar to warnings to deescalate conflicts.

Remove or Not to Remove

While it might seem that moderators were able to choose freely from a suite of possible moderation actions against rule-violating content or people, there were underlying trade-offs under these actions, and moderators had different prioritization of actions to take. For example, while several studies documented moderators prioritizing warnings over direct punishments such as removal or banning (e.g., Jiang et al., 2019; Skousen et al., 2020), I also saw communities that were less hesitant to employ these harsher sanctions (Seering et al., 2019). Furthermore, with any of these actions, moderators had an additional choice to make: whether or not to provide explanations. These different prioritizations reveal two immediate trade-offs. The first trade-off is one that is well-documented by prior research: too much leniency may expose the community to harm, while too much harshness may stifle the community (Gurzick et al., 2009; Kraut et al., 2011). The second is more subtle: removing violating content or people prevents them from staying in the community, but it also forgoes the opportunity to educate the community about acceptable behavior. Behind different competing choices in these trade-offs around moderation actions are differences in moderators' philosophies and values, which I discuss later in this chapter.

5.4 Trade-offs in Moderation Styles

In addition to moderation actions, trade-offs are also present in how the moderators carry out these actions, which I name moderation styles. These moderation styles resemble "distinctions"

in Grimmelmann's (2015) moderation framework (though the identified styles here do not cover all of them), serving as "adverbs" that describe the actions ("verbs") mentioned in the previous section. In my analysis, I identified three major trade-offs around moderation styles in my corpus: human vs. automated, centralized vs. distributed, and transparent vs opaque.

Human vs. Automated

The trade-off between human and automated moderation refers to whether a moderation action was performed by a human or some type of automated system. It is important to note that current moderation systems are rarely purely human or purely automated, nor did any study in my corpus argue for a move toward either of these extremes. Moderation systems that I saw are always a hybrid of human and automated moderation, but the degrees to which they rely on humans or automation vary.

Arguments for more human moderation most commonly appeared when the moderation decisions were difficult and required a nuanced understanding of contexts:

> Moderators we interviewed were happy to have tools that deal with the most obviously unwanted content, such as links to malware or pornography, but they have a strong preference to make the hard decisions themselves. (Seering et al., 2019, p. 14)

One example of such unwanted content that was not obvious was memes, which derives their meanings from multiple layers of contexts (J. A. Jiang et al., 2018). Therefore, in response to Facebook's image recognition tool, Procházka (2019) questioned technology's ability to understand

memes whose meanings were fluid and context-dependent, and argued for the necessity of human moderation of them.

Another thread of cases that warranted more human moderation was community-building. Seering et al. (2019) found that negotiation of acceptable and unacceptable behaviors was critical for community growth, and such negotiation necessarily requires human involvement.

However, negotiation of community norms can take on many shapes and forms. Jhaver et al. (2019) focused on a particular one of them: providing removal explanations, and argued that subpar explanations can have detrimental effects to the community:

> In cases where the removal reasons are unclear, human moderators should continue to provide such explanations. … We expect that inaccurate removal explanations are likely to increase resentment among the moderated users rather than improve their attitudes about the community. (Jhaver, Bruckman, et al., 2019, pp. 22–23)

Despite the benefits of human moderation, moderation research also described the pressing need for automated moderation. As online communities quickly grew into sizes that humans could not reasonably handle (e.g, millions of users), automated moderation provided a solution for moderation at scale (Chandrasekharan et al., 2019). In addition to the ability to moderate large volumes of content, speed was also an advantage of automated moderation that humans struggled to achieve. As the prerequisite for human moderation was that a human had to be online and see the potentially violating content, automated moderation triumphed in timeliness by offering 24/7 monitoring (Jhaver, Birman, et al., 2019). Beyond scalability, Jhaver et al. (2019) also noted that

automated moderation offered a high level of consistency, since moderation rules were hard-coded into the automated systems. However, such consistency presented a tradeoff when facing the unique adaptability to contexts offered by humans, which Jhaver et al. also acknowledged.

Humans' ability to understand nuanced contexts became important in complex, high-stake situations such as when distinguishing hate speech from newsworthiness (Caplan, 2018), where the line between violating and non-violating was critical but blurry. Prior research extensively documented the trade-off between automated tools' ability to handle massive scale of content and human's ability to tell the subtle difference between whether certain content is violating rules (i.e. to reduce false positives), in various cases such as pro-eating disorder communities (Chancellor et al., 2017, 2018), crowdsourced blocklists (Jhaver et al., 2018), copyright infringement detection (Gray & Suzor, 2020), and even country-wide ethnic violence (Jhaver, Birman, et al., 2019). Chancellor et al. (2017) specifically pointed out that automated tools could magnify any errors they made, as well as the remedy required to correct these errors, precisely because of their ability to scale.

To summarize, studies reveal benefits in both human and automated moderation: Humans are capable of handling complex nuances, while automated systems offer the kind of moderation required by the massive scale of today's online community. However, these very benefits can become drawbacks in different situations, and the trade-offs between human and automated moderation remain a persistent challenge to content moderation.

Centralized vs. Distributed

The trade-off between centralized and distributed moderation refers to whether moderation decisions are made by designated moderators or regular users and community members. Similar to human vs automated moderation, the configuration of centralized vs. distributed moderation is

often a hybrid one in today's online communities, landing somewhere between purely centralized and distributed. For example, Facebook has centralized moderation teams around the world to enforce their community guidelines, as well as volunteer moderators in Facebook Groups to make their own rules and enforce their own moderation (Gillespie, 2018a). Likewise, Reddit also has platform-wide moderators as well as volunteer moderators in individual subreddits that form the moderation system on Reddit that we see today (Fiesler et al., 2018). Even within the premise of a single subreddit, many subreddits also allow regular members to contribute to moderation decisions such as rule making in addition to the moderators. Furthermore, Reddit users also have the ability to upvote or downvote posts, which impacts the visibility of these posts (Fiesler et al., 2018).

Many papers pointed out drawbacks of distributed moderation that indicate the advantages of a centralized fashion. The arguments against distributed moderation focused on the lack of expertise from regular users, as well as their personal biases which made them incapable of making decisions representative of the community ideal:

> r/AskHistorians moderators described a variety of reasons why they opposed using the karma system as an indication of quality. First, the majority of those who upvote responses do not have the requisite expertise to evaluate quality; second, voting reflects user bias; and third, earlier comments tend to receive more upvotes, regardless of quality. (Gilbert, 2020, p. 15)

These drawbacks of distributed moderation suggested that centralized moderation would be more consistent, standardized, and made by qualified experts. Some participants in Fan and Zhang's (2020) digital jury experiment expressed a similar lack of confidence in the quality of user input.

Furthermore, Duguay et al. (2018) found that distributed moderation could harm minority users disproportionately:

> Co-moderation works against minority user groups on two levels. First, the majority of users on such a mainstream platform as Instagram are statistically more likely to be heterosexual and may have difficulty understanding the aims and culturally specific aesthetics of queer women's photos. Secondly, those who are compelled to flag others' photos do so because they feel strongly about the content, usually because they are offended by its violation of their personal norms, which may be sexist or homophobic. (Duguay et al., 2018, p. 18)

Here, Duguay et al. suggested that decisions from distributed moderation could favor majority norms against marginalized groups, a finding echoed by Park et al. (2016) in pointing out the "undesired popularity bias" in crowdsourced moderation of news comments.

However, distributed moderation also has desirable advantages. I saw many cases where users had higher confidence in distributed moderation over centralized moderation (e.g., Ehrett, 2016; Seering et al., 2019), with one study (Draper, 2019) specifically arguing that distributed deliberation practices could foster a positive digital environment. In their study, Fan and Zhang (2020) found that compared to distributed moderation, centralized moderation was less democratically legitimate in the framework of procedural justice, characterized by a lack of accountability to the public.

Furthermore, since centralized moderation converged moderation to a small team of moderators, they had to "spend countless hours in order to maintain the community" (Chandrasekharan et al., 2018), which suggested distributed moderation's potential ability to diffuse

moderators' workload. The ability to reduce workload, however, was at odds with the desire of expertise in moderation, which was the major advantage of centralized moderation and typically only the moderators possessed. Lampe and Resnick (Lampe & Resnick, 2004), in one of earliest studies of content moderation, summarized this inevitable trade-off between improving efficiency and seeking expertise:

> These findings highlight tensions among timeliness, accuracy, limiting the influence of individual moderators, and minimizing the effort required of individual moderators. We believe any system of distributed moderation will eventually have to make tradeoffs among these goals. (Lampe & Resnick, 2004, p. 8)

In addition to the moderation work itself, the expertise desired in centralized moderation and the public accountability desired in distributed moderation also highlight another trade-off: is the credibility derived from the experts or that derived from the public more desirable? Kayhan et al. (2013) rightfully pointed out this trade-off in perceived credibility, and came to the conclusion: It depends.

> [G]overnance credibility is a contextual variable that varies from one situation to the next. Governance mechanisms implemented in two different organizations may not be equally credible if the governors are different. In a given context, expert-governance may be perceived as being more credible than community-governance if users trust the experts more than the community members (or vice versa). (Kayhan & Bhattacherjee, 2013, p. 75)

In summary, I found that the trade-off between centralized and distributed moderation was one that revolved around perceived expertise, efficiency, and credibility. Just like the case of human vs. automated moderation, my analysis indicates that the centralized vs distributed trade-off may be inevitable.

Transparent vs. Opaque

The trade-off between transparent and opaque moderation is prominent in my dataset. While this trade-off is similar to the distinction of transparently vs. secretly in Grimmelmann's (2015) moderation framework, Grimmelmann's distinction focuses more on whether the fact that some kind of moderation had happened is explicit and public. However, the distinction between transparency and opacity here in my dataset focuses more on whether explanations are provided with sanctions, and the visibility of the act of moderation is less of a concern.

I saw an undeniable push for transparency in my analysis, with ample discussion of the benefits of providing explanations. Studies found that transparency enhanced legitimacy, perceived consistency (Witt et al., 2019), and accountability (Fan & Zhang, 2020), and could prevent confusion and frustration that breeded the often incorrect folk theories for why certain content was sanctioned (Jhaver, Appling, et al., 2019; J. A. Jiang et al., 2019; Suzor et al., 2019; West, 2018). Providing explanations also helped community members adhere to norms and improve their future behaviors (Jhaver, Bruckman, et al., 2019; Tyler et al., 2019), and educated users about community rules (Jhaver, Appling, et al., 2019).

Despite a multitude of benefits of being transparent, I also saw valid reasons for not providing explanations. Many studies (Jhaver, Birman, et al., 2019; Jhaver et al., 2018; Juneja et al.,

2020) reported that providing explanations of actions by automated moderation tools enabled malicious actors to game the rules:

> We found that moderators do not reveal the details of exactly how AutoMod[erator] works to their users. … Our participants told us 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. (Jhaver, Birman, et al., 2019, p. 21)

Chancellor et al. (2016) explored such circumvention of hard-coded rules in detail through a case study of how pro-eating disorder communities used lexical variation to avoid hashtag-based moderation on Instagram, which did not even publicize how it moderated hashtags. The need to prevent rule circumvention extended beyond tool configuration to community rule making itself: Many moderators chose to phrase their rules vaguely and broadly so that they could have the necessary interpretative flexibility when it came the time to enforce these rules (J. A. Jiang et al., 2019; Juneja et al., 2020).

Explanations provided by humans had different problems. Contrary to recent findings, Petrič and Petrovčič (2014) found that providing explanations did not increase users' sense of community. Furthermore, Seering et al. (2019) found that transparency could be a source of conflict within communities, because community members often would not notice unannounced moderation decisions. Possible disagreements and conflicts resulting from transparency could escalate to harms against moderators, as Gilbert (2020) suggested in her study of r/AskHistorians:

> While the stickied [explanation] comment may have reduced the total number of questions and comments than the question would have received without the stickied comment, it did not solve the problem entirely and resulted in additional emotional labor as users responded to the stickied post with insults. (Gilbert, 2020, p. 22)

Several other studies echoed the emotional labor associated with moderation (e.g., Dosono & Semaan, 2019; Wohn, 2019), but the physical labor as well. Providing explanations is a nontrivial amount of work. Jhaver et al. (2019) advocated the use of automated tools to provide explanations to handle the enormous traffic that online communities often experience today. However, as I mentioned previously, automated tools have the potential to magnify their errors, and tools mistakenly providing the wrong explanations could exacerbate the conflict and hostility toward moderators.

The trade-off between transparency and opacity is difficult, with no benefits of one side clearly outweighing those of the other. In an in-depth study of Reddit's moderation transparency, Juneja et al. (2020) made this trade-off prominent by showing that their moderator participants had divided opinions on almost every issue related to moderation transparency, including whether or not to make removals obvious, to provide explanations for sanctions, to share details of AutoModerator implementations, and to make moderation logs public, for the same reasons I discuss above. The tug of war between improving behaviors, legitimacy, and accountability, and preventing rule circumvention, conflict, and attack toward moderators remained a subtle balance to achieve in content moderation.

Overall, these three trade-offs in moderation styles, together with the trade-offs in moderation actions that I discussed in Section 5.3, reflect deeper decision making rationales in content moderation, which I discuss in the next section.

5.5 Trade-offs in Moderation Philosophies

The moderation actions and styles above reflect moderators' varying moderation philosophies, which are prioritizations of competing needs that led to the actions and styles that the moderators chose to employ. In my dataset, I identified three major trade-offs in moderation philosophies: nurturing vs. punishing, efficiency vs. quality of moderation, and level of activity vs. quality of contribution.

Nurturing vs. Punishing

Nurturing and punishing both aim to create a positive online environment, but reflect different ideals in moderation's purposes, which Ruckenstein and Turunen (2020) conceptualized as "the logic of choice" and "the logic of care." Nurturing takes an educational approach that aims to improve or reform community members' behavior, while punishing focuses on removing the rule-violating content from the community, and making sure that the rule-violating person receives consequences for their behavior.

I saw nurturing typically associated with less harsh and more educational actions like providing warnings, offering explanations, and actively diffusing conflicts in the community. Seering et al. (2019) noted that moderators who took a nurturing approach saw misbehaviors as something to be reformed rather than to be eliminated:

> Rather than seeing misbehavior as something that could be "cleaned up" by algorithms or bans, many moderators choose to engage personally during incidents to set an example for future interactions. (Seering et al., 2019, p. 2)

A reformative approach can be desirable especially because not all misbehaviors come from malicious perpetrators who intentionally disrupt communities. Jhaver et al. (2019) found that some people broke rules simply because they unintentionally overlooked the rules, and argued that it was worthwhile to nurture these sincere users by offering explanations so as to not drive them away. Furthermore, as I discussed in the transparent vs. opaque trade-off, providing explanations to educate users could improve their behavior as well as their perceptions of content moderation in their communities. These benefits had prompted researchers to argue for a nurturing rather than punitive approach in content moderation (Jhaver, Appling, et al., 2019; West, 2018).

However, punishing can also be valuable in community maintenance. While arguing for a general nurturing approach to moderation, Jhaver et al. (2019) also highlighted the necessity of punishment:

> We 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. (Jhaver, Appling, et al., 2019, p. 26)

While suggesting differential treatments of rule violation between well-meaning and malicious people, Jhaver et al. rightfully pointed out the limitation in human moderation resources—providing detailed, customized nurturing requires human work, a point I have reiterated in discussing the transparency vs. opacity trade-off. Furthermore, more human resources invested in nurturing meant less in punitive actions such as removal, which was necessary to remove harmful content to prevent them from overwhelming legitimate content. Einwiller and Kim (2020), through a study of online content providers in four countries, extended Jhaver et al.'s volunteer moderation-based arguments to commercially-moderated platforms, highlighting the heightened difficulty of a nurturing approach when the scale was much larger:

> [Interviewees] stated that decisively pointing out publicly where and why comments violated the policy and referring to the respective policy could help educate the poster and those observing. When the volume of [harmful online content] is large, however, doing so is often impossible. It is also a challenge to do this when a user is clearly trolling or posts are severely harming others so that they have to be removed immediately. (Einwiller & Kim, 2020, p. 198)

Einwiller and Kim (2020) identified severity as another key reason for taking a punitive approach to prevent exposing platform users to harm. In my analysis of platforms' community guidelines in the last chapter, I found that platform moderation had to face a wide range of harmful content, from insensitive jokes to coordination of mass murder. The latter obviously requires immediate removal and possibly an account ban, rather than a kind, educational message saying that mass murder does not contribute to a positive online environment. The severity-based moderation philosophy applied not only to platforms, but to smaller communities as well (Blackwell et al., 2019;

J. A. Jiang et al., 2019; Seering et al., 2019). Therefore, the configuration in the nurturing vs. punishing trade-off, like in all other trade-offs, is a hybrid one in practice, with differing tendencies toward one or the other depending on the specific community context.

Level of Activity vs. Quality of Contributions

The trade-off between level of activity and quality of contributions is related to content in the community. It represents competing desires of a large amount of traffic in a community (e.g., a large number of members, a high amount of daily posts), and high quality contributions in the community (e.g., correct categorization, minimum low-effort posting[1]).

The trade-off between level of activity and quality of contribution relates to how strictly moderators enforce the community rules, which represents a trade-off that I have discussed in moderation actions: Loose moderation retains community members but may also retain low quality or even harmful content, whereas strict moderation promotes high quality content but may stifle the community (Gurzick et al., 2009; Kraut et al., 2011). Srinivasan et al. (2019), for example, concluded that strict moderation through removal contributed to a high quality of community content, but also acknowledged the possibility that authors of the moderated posts might get discouraged and leave the community. Furthermore, research (Jhaver, Bruckman, et al., 2019) found that providing explanations, the more nurturing and less punitive approach than mere removal, also had the potential to alienate users and drive them away, noting "moderators may need to consider whether having high traffic is more important to them than having quality content on their community."

[1] Often called "shitposting" in online communities.

The battle between traffic and quality was also one that community members realized. Jhaver et al. (2019) found that community members would intentionally break rules that ensure clean organization of community content, which in their case, was a rule that mandated that questions are only posted in designated threads. While community members acknowledged the purpose and necessity of that rule, they believed that it made individual questions invisible and "stifled community interactions," and chose to break the rule with speculations that their posts would subsequently be removed.

Here, it is clear that making their own questions visible was more important to these community members, and the potential benefits outweighed the risks of breaking the rule. However, considering the scale of today's online community, having questions scattered in the community without a centralized repository (e.g., a question thread) may overwhelm other members to the community. Similarly, members of the r/NoSleep subreddit noted that while strict regulation helped them survive the surge of newcomers as a result of becoming one of the default subreddits, it also deprived old members of the kind of freedom they used to enjoy (Kiene et al., 2016). Therefore, having to face different types of community members, the answer to the level of activity vs. quality of contribution trade-off may not be obvious to the moderators.

Efficiency vs. Quality of Moderation

While quality of community contribution was an important consideration, so was another kind of quality—the quality of moderation. The trade-off between efficiency and quality represent two competing characteristics of content moderation work. On the one hand, moderation needs to be efficient in order to monitor and handle content in a timely manner. On the other hand, moderation also needs to fulfill goals related to quality, which converge to the central goal of making

sure all content receives appropriate treatment. There is a reason for the overly-broad definition of quality: Later I will show that the meaning of quality is complex, and consists of multiple, sometimes competing factors.

The reason for the need for efficiency is straightforward: undesirable contents should not stay up for too long. Undesirable contents range from unuseful to harmful, and the longer they exist, the more impact they have on the community. Moderation research from the earliest time has expressed the desire for efficiency (Lampe & Resnick, 2004; Wang et al., 2014), which became one of the primary reasons for the widespread use of automated moderation tools (Jhaver, Birman, et al., 2019). Minimizing delay in moderation has become even more important as online communities gain variety. The most prominent examples are communities with real-time interactions. Seering et al. (2019) noted that on the live streaming platform Twitch, "due to the synchronous nature of conversations … moderation decisions need[ed] to happen immediately." In voice-based communities on Discord, interactions were not only in real-time but also ephemeral. Therefore, unless moderation could happen with virtually no delay, moderators needed to seek evidence of rule breaking to make sure it even happened, a problem Discord moderators were constantly facing (J. A. Jiang et al., 2019). Furthermore, the need for efficiency did not only exist for identifying the misbehavior, but also for deciding what actions to take on the misbehavior:

> However, participants also described aspects they did not like about deliberation. Eight people mentioned lower efficacy. One user identified a trade-off between efficacy and richer user input. (Fan & Zhang, 2020, p. 8)

In Fan and Zhang's (2020) digital jury experiment where they recruited users to serve as "jurors" that make moderation decisions, they found that deliberation between the jurors delayed the

moderation decision in sacrifice to careful, democratic decision making. The digital jury example is exemplary of the trade-off between efficiency and quality in question, with quality represented by "richer user input."

The meaning of moderation quality, however, is more complex. Different studies conceptualized "quality" differently, as already shown by previous sections. For example, Fan and Zhang (2020) considered "quality" to be democratic legitimacy and accountability. Schoenebeck et al. (2020) argued that "quality" should be customized moderation that did not fail some people while privileging others. Jhaver et al. (2019) believed "quality" of moderation to be minimal incorrect decisions (i.e., "false negatives" and "false positives"), though the concept of "correct" might be just as complex as "quality." In a tricky case of r/AskHistorian where moderators had to choose between directly removing a post and explaining why that post was subtly harmful, Gilbert et al. (2020) presented an example of almost complete surrender of efficiency for the pursuit of high-quality moderation, where a moderator biked to a nearby, paywalled library to find answers to a community member's question.

While these examples are by no means comprehensive, all of them require human deliberation and thus sacrifice of efficiency to various extents. As the trade-off between efficiency and quality pertains to every moving part in what I have described in the previous sections about moderation actions and styles, it might be the case that decisions in different parts will compete with each other and impact the overall efficiency vs. quality trade-off as a whole.

Overall, these three trade-offs in moderation philosophies presented in this section represent competing desires for the purpose of moderation, the process of moderation, and the community

content shaped by moderation. These subtle decisions in philosophies reflect values that different stakeholders in online communities hold, which I discuss in the next section.

5.6 Trade-offs in Moderation Values

So far, I have discussed many trade-offs in moderation actions, styles, and philosophies. These trade-offs show competing needs that are all legitimate, have pros and cons, and do not have clear, "right" answers. However, facing these trade-offs, moderators must make decisions, and I found that these decisions were impacted by trade-offs in the values that they might hold. In my dataset, I identified three facets in the trade-offs in moderation values: moderator identities, community identities, and competing stakeholders.

Moderator Identities

Moderator identities are what moderators see themselves as in their communities, such as governors, teachers, and gardeners, to give a few examples. Prior research (e.g., Wohn, 2019) often referred to these identities as roles characterized by designated tasks, but here I use the term "identity" to emphasize moderators' self-perceived high-level responsibilities that transcend specific tasks. The differences in moderator identities have been a prominent theme since the earliest of moderation research in my corpus:

> One admin saw his role as being particularly centred on careful management of people in "keeping the peace" and maximising the potential of others, while another saw his role as being more based around the filtering of discussions and the group pool. (Holmes & Cox, 2011, p. 12)

In their study of Flickr administrators, Holmes and Cox (2011) found moderator identities that correspond to the nurturing vs. punishing trade-off in moderation philosophies. As online communities evolve, the perceived identities also start to vary more. For example, Matias (2019) listed a range of identities that his participants self-identified with, along with different corresponding duties. For example, dictators "make all the decisions," janitors "clean up," and martyrs "give hell to anyone who dared to...threaten [their] communities." Similarly, Seering et al. (2019) in their study of 56 moderators also found several identities, including arbiters, community managers, role models, etc.

While I do not get into the detail of the subtle differences and overlaps between the identities listed here (which deserves its own research), moderators used them to justify the moderation decisions they had made. While prior research has not always made explicit connections between these identities and philosophies, styles, and actions, it is reasonable to speculate that moderators who identify as arbiters would prefer to adopt centralized moderation, or those who identify as curators would care about the quality of contributions more than the level of community activity. Overall, taking up a certain identity means to serve certain responsibilities and purposes, and to take actions accordingly (Wohn, 2019).

Wohn also pointed out that moderator identities were not mutually exclusive. The co-presence of these different, sometimes competing identities showed that there was a need for many of them—for example, moderation may need to be nurturing and punishing, instead of nurturing or punishing. Another line of research on social roles (e.g., Yang et al., 2019) also echoed these simultaneously existing identities. Gurzick et al. (2009) described how moderators were aware of the need to balance identities, and that moderators "debated the proper role that they should take and

negotiated the amount of activity that would be reasonable." The negotiation of these identities shows that making decisions in the trade-offs in actions, styles, and philosophies may extend beyond their own pros and cons, to a deliberation of value differences.

Community Identities

In addition to how moderators see themselves, how moderators see their communities also has an impact on how they moderate them. I call communities' self-conception of what kind of communities they are as "community identities." The perceived identity of a community determines who and what is welcome or unwelcome in the community, and what purpose the community is supposed to serve.

A prominent trade-off in community identities is that between, as Gibson (2019) named, "free speech," and "safe spaces." The former referred to online spaces that promote free expression of opinions, while the latter emphasized mitigating potential harm that speech could cause. Gibson (2019) found that compared to "free speech" spaces, in "safe spaces" moderators removed significantly more content, indicating a punitive tendency that focused more on the quality of community content (in this case, content that did not harm marginalized communities). Like Gibson, other research (Grover & Mark, 2019; Phadke & Mitra, 2020) also revealed these two often competing identities, highlighting that it is a difficult trade-off to balance:

> As political and ideological stratification in society continues to grow, and online communities focused on ideological commitments become more numerous, moderators of online platforms … face difficult challenges in how to balance the right to free expression, with broader concerns of public safety and wellbeing. (Grover & Mark, 2019, p. 203)

Trade-offs in community identities also existed for communities committed to certain topics, where moderators struggled to balance competing conceptualizations of the topic. For example, in r/Paleo, a subreddit for the paleo diet[1], moderators struggled to maintain a balance between a consistent community conception of paleo diet and individualized understandings of what paleo diet is:

> Paleo faces a tension between the need to maintain some kind of coherent concept of the diet while also allowing flexibility for adherents to pursue a diet that accounts for individual differences. One way of negotiating this tension comes through the rules of the subreddit. One of the only rules that r/paleo moderators actively enforce is not to "[a]ct like your One True Paleo™ is the be-all, end-all and is perfect for every human on Earth." (Squirrell, 2019, p. 1919)

Here, the moderators did not decide on one particular identity to pursue as a community, but simply required that members keep an open mind toward all versions of the paleo diet.

The differences in how certain topics are conceptualized also exists in research of these communities, with pro-eating disorder (pro-ED) communities the most prominent. A long line of research (Chancellor et al., 2017, 2018; Chancellor, Pater, et al., 2016; Feuston et al., 2020; Gerrard, 2018) on moderating pro-ED communities shows a clear trajectory of how pro-ED communities are viewed: from communities that promote eating disorders as a legitimate lifestyle, to those that support and help people with eating disorders. The co-presence of competing

[1] Wikipedia explains paleo diet as "a modern fad diet consisting of foods thought to mirror those eaten during the Paleolithic era."

conceptualizations meant that the same content could be treated differently due to (1) how they were perceived, and (2) whether that perception matched the community identity:

> Harm reduction provides resources for individuals who have an eating disorder, but cannot or will not recover, to stay safe and informed. Despite benefits, harm reduction resources are treated differently across eating disorder spaces online. While some communities freely permit them, others, such as one of the active subreddits in our digital ethnography, have moderation teams dedicated to removing posts related to tips or advice and carefully overseeing content related to harm reduction. (Feuston et al., 2020, p. 16)

Feuston et al. (2020) argued that content moderation should consider the full complexity of marginalized experiences such as eating disorders, and not cast negative stereotypes on content like harm reduction that might help those in need.

While Feuston et al. provided an example of how fulfilling stereotypical community identities could be harmful, Gilbert (2020) further complicated the issue by demonstrating how fulfilling seemingly innocent identities could also cause unintentional harm. In the same example I discussed in the efficiency vs. quality of moderation trade-off, where an r/AskHistorians member posted a question about the background of a historical photo featuring naked women in the military, fulfilling the community identity became at odds with the need of being contextually sensitive:

> In circumstances in which biased or insensitive questions are asked, moderators are tasked with making the decision to let the question stand or remove it, and experts with the decision to respond to the question or ignore it. … During our

> interview, moderator, Mark Evans described deliberating whether or not to remove the question: "We had a discussion about removing it because the pictures are incredibly … exploitative … And we just felt so shitty as moderators, because here was our community, which is meant to be giving people answers about the past, but what it's doing is providing Redditors with porn. And that's what it ended up doing. And that's why people have ended up looking at it and it's become a platform for these poor women to become humiliated again, like 80 years after the event. Again." (Gilbert, 2020, pp. 11–12)

As Gilbert later pointed out, the issue of whether or not to provide people with answers about an exploitative past raised questions about trade-offs between centralized and distributed moderation, as well as "free speech" and "safe spaces." While prior research argued that community identity might not be as salient in the moderation of platforms due to the lack of strong ties (Fiesler & Bruckman, 2019), the discrepancy of perceived platform identity could still be a source of conflict, like when Yelp users left one-star reviews for a merchant that employed someone who had contentious political beliefs on immigration, many of which Yelp removed (Medeiros, 2019). While Yelp intended the reviews to be about the commercial services of merchants, the users found them as "symbolically significant means of signaling social disapproval." Medeiros (2019) characterized the unintended use of reviews as "a genuinely vexing moderation challenge for Yelp, suggesting a limit to the site's ability to enforce rules that dichotomize political and commercial content."

In both examples above, core to the problems is the different prioritization of community identities across different stakeholders. I explore the impacts of different stakeholders in the next section.

Competing Stakeholders

Moderation is often expected to satisfy multiple stakeholders and their often different needs, which presents a difficult task for moderators who often have to make decisions that serve some over others. Matias (2019), for example, summarized volunteer moderators' work of serving different stakeholders as their "civic labor":

> This "civic labor" requires moderators to serve three masters with whom they negotiate the idea of moderation: the platform, reddit participants, and other moderators. Moderators differ in the pressure they receive from these parties and the weight they give them. Some face further stakeholders outside the platform. Yet attempts to make sense of moderation by focusing on any one of these relationships can bring the other actors out of focus. (Matias, 2019, p. 8)

While I have shown in previous sections the impact of community members, and other moderators, platforms are also a significant factor. Volunteer moderators' power cannot reach beyond the purview of the platform where their communities are hosted, and consequences could be severe when negotiations with platforms fail. One such example is the Reddit blackout, where many moderators shut down their communities in response to Reddit's dismissal of an employee who routinely offered support to volunteer moderators. Matias (2016a) showed that such protest against the platform was still a negotiation among moderators, users, and the platform:

> Reddit employees played a key role in these negotiations [with Reddit]. … Across subreddits of all sizes, relations among moderators were also associated with participation in the blackout. … Community members also played an important

> role in action against the platform by pressuring moderators to join the blackout, discussing and voting in decisions, and sometimes even punishing moderators who disagreed. (Matias, 2016a, pp. 1146–1147)

Like volunteer moderation, commercial moderation faces the same trade-off between multiple stakeholders. The common factor was users—for example, differently politically affiliated users also perceived content moderation differently (Hua et al., 2020; Shen & Rose, 2019). Schoenebeck et al. (2020) also found that people with different backgrounds had significantly different preferences for the kinds of remedy social media sites could offer for online harassment.

However, platform moderation also needed to satisfy a new set of stakeholders. First, unsurprisingly, platforms have to operate under the requirement of local law, which often ban severely harmful content on a statutory level such as child pornography and terrorism (Einwiller & Kim, 2020; Gillespie, 2018b; Zeng et al., 2017). However, these content may still provide value to someone else:

> Often disturbing, graphic, and controversial, human rights-related media like the Werfalli and Syrian war videos pose a dilemma for platforms hosting them, involving difficult tradeoffs between their perceived social value and their possible harms. (Banchik, 2020, p. 2)

Banchik (2020) found that even graphically abusive content may prove to be valuable documentations to various human rights workers, adding that:

> Practitioners I spoke with expressed added concern that biased or merely ill-informed human reviewers "without the requisite knowledge" would decide the

> fate of vital documentation. Moreover, most practitioners did not blame platforms alone for the removal of content, but instead saw the topography of takedowns as far more complex. (Banchik, 2020, p. 7)

Platforms are also aware of the complexity of harmful content given their potential public value. Facebook's Community Standards (Facebook, n.d.), for example, states:

> In some cases, we allow content for public awareness which would otherwise go against our Community Standards—if it is newsworthy and in the public interest. We do this only after weighing the public interest value against the risk of harm and we look to international human rights standards to make these judgments.

However, Facebook's decision to not remove some violence-inciting message on the same ground provoked heated debate among users and various experts (Shieber, 2020).

Furthermore, for platform designers, the fundamental need to moderate content for users becomes a trade-off to consider with the psychological health of moderators. Both academic research (Karunakaran & Ramakrishan, 2019; Luo et al., 2020; Riedl et al., 2020) and journalistic coverage (Newton, 2019) revealed the emotional impact of moderating disturbing content. As platform technologies evolve into new forms like live-streaming video, produced content are more likely to provoke intensified emotional reactions, and therefore what is asked from moderators, both logistically and emotionally, can also escalate (Luo et al., 2020).

Above are only some of the examples of the full complexity of content moderation in a multi-stakeholder environment. Realization of the needs of multiple stakeholders has prompted many studies to call out against a one-size-fits-all approach to content moderation (Blackwell et al.,

2017; Gallagher & Savage, 2016; J. A. Jiang et al., 2019; Schoenebeck et al., 2020). However, as desirable as customized moderation might be, it may not be entirely feasible due to constraints in human and technological resources. Then, whose needs are prioritized, and what downstream impacts it has on various trade-offs, are critical problems to consider in content moderation.

5.7 How Different Stakeholders Can Use the Framework

My framework offers a different way to examine content moderation, one that posits trade-offs in the front and center. As an example, Seering et al.'s (2019) findings on the differences in actions taken by moderators toward misbehaviors indicate that values impact moderator actions. However, if we examine their findings through the lens of my framework, we can reveal several additional research questions related to the trade-offs that could have happened: While communities with more laissez-faire ideologies use less equivalents of bans than those intended to be "safe spaces," what prompted the communities to side with certain ideologies over others? Do moderators' perceived identities differ between Reddit and Facebook, and does that have an impact on differences in the level of reliance on automated tools? These are only a few examples of the questions we can ask from the application of my framework, and Seering et al.'s (2019) speculation of the answers to the latter question testified to the value of my framework—"The difference [in the preference of automated tools] likely results from the importance of continuously evolving community values in decisions made by moderators." Answers to these questions will offer a deep, rich understanding of the inner-working of content moderation from a new angle.

The above example is only one way **researchers** of content moderation can use this framework as an analytical tool in their own research. For example, a researcher can use my framework to identify key trade-offs in moderators' decision-making, investigate why moderators

took certain actions instead of other actions they could have taken, and trace back to their philosophies and values behind these decisions. Furthermore, researchers can also use my framework to identify potential value tensions behind certain philosophies, and potential caveats of recommendations they might make. For example, when recommending that the moderation of a community or a platform should be more transparent, what are the potential stakeholder tensions that may prevent it from doing so? How can it resolve such tensions to get closer to the researchers' ideal?

Designers of content moderation can use my framework as a heuristic for their design, either improve an existing content moderation system, or to build a new one. Designers who wish to improve an existing moderation system can use my framework to identify key decision points that moderators may struggle with and to be critically aware of the trade-offs and tensions involved. While their designs may inevitably favor one side of a trade-off, designers can consciously find their ideal balance in the trade-off so that their designs can be more considerate of the other side. Similar to the case of researchers, some trade-offs may not be applicable or salient to some communities or platforms. While designers should focus on the trade-offs as appropriate, with my framework they can also consider making some previously invisible trade-offs more salient as a potential form of improvement.

Designers who wish to build new content moderation systems can use my framework as a guide to support moderators in key decision points. For example, designers may consider explicitly showing the available actions and decisions to moderators as trade-offs instead of a simple listing, as well as the potential consequences of making different decisions. Designers can present these trade-

offs not only in manuals or training materials, but also in the interface of moderators' day-to-day work, so that moderators can be more informed when making decisions.

Moderators may also benefit from my framework as a way to encourage reflexivity in their own work. For example, my framework will allow moderators to realize that when they make a decision on doing something, they are also making decisions on not doing something else. Therefore, moderators will be able to make more conscious trade-offs in their work, and have elaborate justifications for past decisions that may be valuable for revising or improving their workflow.

Finally, **users**, or people who are moderated, may find values from my framework when participating in content moderation in various ways. A key element in content moderation, users will be able to learn the full complexity of moderation from the trade-off-centered framework, and therefore be more informed when disputing moderation decisions, contributing to rule making, or engaging in conversations about content moderation in general.

6
conclusion

"The truth is, we wish platforms could moderate away the offensive and the cruel. We wish they could answer these hard questions for us and let us get on with the fun of sharing jokes, talking politics, and keeping up with those we care about. But these are the fundamental and, perhaps, unresolvable tensions of social and public life. Platforms, along with users, should take on this greater responsibility. But it is a responsibility that requires attending to these unresolvable tensions, acknowledging and staying with them—not just trying to sweep them away."

—Tarleton Gillespie, in Custodians of the Internet

In the preceding chapters, I have examined different stakeholders in online content moderation, revealing their varied interests and perspectives, the problems that could arise when their needs are overlooked, and proposed a trade-off-centered model of understanding online content moderation. In this chapter, I first discuss lessons learned from and provide recommendations for moderating new technologies and diverse, globally users. Then I discuss what it means to make choices among competing options when content moderation is expected to serve various technologies and people. These considerations of different stakeholders challenge existing assumptions, practices, and designs in content moderation that often attempt to fit everything into one mold when facing the problem of scale. By highlighting the importance of embracing stakeholders and deliberately weighing their needs, my research marks a multi-stakeholder future necessary for content moderation.

6.1 Moderating New Technologies

My investigation into the moderation of voice-based communities on Discord in Chapter 3 is a demonstrative example of how existing notions of moderation can fail in communities with emergent technologies, and more broadly, how overlooking unique characteristics of different communities can result in serious consequences in content moderation. Using voice moderation as a case study of moderating new technologies, I first provide specific recommendations given the unique affordances of voice, and then discuss how the lessons in moderating voice can point to broader conceptual recommendations for moderating emergent technologies in online communities.

Recommendations for Moderating Voice

The findings of my empirical study point to several implications specifically for the design of voice moderation tools. First, many rule violations that moderators identified in voice channels may

be preventable using automated systems. Moderators told me that slurs and hate speech were common in voice channels. It is tempting to consider the use of automated systems to detect this type of speech, as many moderators already did in text-based communities (Seering et al., 2019). However, my findings also show that the intonation can be nuanced, and moderators took context into account and made moderation decisions on a case-by-case basis—which would be challenging for an automated system to do. Furthermore, these systems may unfairly punish those who do not speak English in the same way that these voice recognition systems were trained (Harwell, 2018; Paul, 2017). As Grimmelmann (2015) pointed out, any moderation action can nudge the community norms in an unpredictable direction, and a blanket ban of what automated systems deem inappropriate may do more harm than good for a community. This caveat applies to any moderated online community, but the potential harm may be more pronounced in communities in which the underlying rules and norms are evolving or unclear, as is the case in many Discord voice channels.

To prevent disruptive noise, for example, platform or system designers can design systems that detect volumes that may be uncomfortable for humans. An intuitive implementation of this system would be to automatically mute accounts that are too loud, but my findings suggest that loudness can be a result of misconfigured hardware. Therefore, while temporary muting would still be necessary, the system may also want to prompt loud accounts to check their hardware settings. Similarly, to mitigate music queue disruptions caused by lengthy audio, a system could alert moderators of audio that exceeds a certain length threshold, and let them determine whether it is a rule violation. Both of these design recommendations aim to prevent rule violation preemptively, rather than reactively.

To address the major need for moderators to acquire evidence in voice channels, platforms may want to incorporate recording functionality within moderator tools. While there are already third-party applications that are able to record voice channels on Discord, such as MEE6, it may still be better for platforms to have control over such features. Third party applications may pose privacy risks to users, but if implemented within the platform, platforms could both take measures to protect the recorded data, and make sure that users are informed. Furthermore, without access to the platform directly, third party applications typically only record audio, which, as my findings suggest, is not sufficient when moderators need to connect distinct voices to user accounts. Therefore, platforms may also need to generate video files, or audio files with metadata, that show who is speaking (or not) at any time.

However, I also recognize that the design recommendations above largely require some type of automated system to listen in the voice channels at all times, which may raise privacy concerns among community members. This issue, together with my prior discussion of "incognito recording" in Chapter 3, steps into the legal and policy realm of recording conversations. One-party consent recording—recording by a participant in the conversation without other parties' consent—is against the law in eleven states in the U.S. and some other countries ("Telephone Call Recording Laws," 2019). Discord's Terms of Service does not have rules about recording, nor any other behavior specific to voice. However, it does prohibit users from "engaging in conduct that is fraudulent or illegal," and require users "to comply with all local rules and laws regarding your use of the Service, including as it concerns online conduct and acceptable content," recognizing that Discord is an international platform. These blanket statements mean that the work of deliberating different

regulations around the world is offloaded to the individual server moderators—something Discord explicitly states in its Community Guidelines (Discord, 2019):

> We do not actively monitor and aren't responsible for any activity or content that is posted; however, we expect server owners, mod[erator]s, and admins to uphold their servers to these guidelines and we may intervene if community guidelines are not upheld.

However, it is not reasonable to assume that all moderators would know all the regulations in the world, particularly since people often have incorrect interpretations of both the law and Terms of Service provisions (Fiesler et al., 2016). The findings in Chapter further show that the complexity of moderation in a global landscape is far more than any volunteer moderator can reasonably handle. One of my participants stated confidently that recording was definitely not against Discord's Terms of Service—though given the complexities of the broader laws that the document nods to, this may or may not be true. Despite these complexities, my findings show that recording could be a desirable solution to the major problem of gathering evidence. Therefore, to prevent volunteer community moderators from bearing legal consequences unknowingly, platforms like Discord could either explicitly acquire consent at the platform level (e.g., in Terms of Service, or through a pop-up dialog when a user joins a voice channel for the first time), or advise individual community moderators to explicitly acquire consent within their communities if they wish to record voice channels.

Stakeholders Using Different Technologies

Voice as a technology has been around for decades, but voice-based communities had only recently became a stakeholder in the context of an online community moderation that has largely focused on text. While the empirical study in Chapter 3 only focuses on moderating voice, online

communities are bound to continue to develop beyond text and voice. Though specific issues may not generalize beyond voice, the types of problems I have identified could appear in other new social technologies where moderation may be necessary. For example, we are already seeing emerging social virtual reality (VR) communities—for example, VRChat, where people interact with virtual avatars and communicate using voice. VRChat shares many characteristics with Discord voice channels: interactions happen in real time and are not recorded. However, the virtual physical presence of VR adds another complication to moderation—for example, there are already reports of users who sexually harass other users physically (Bell, 2018; Blackwell et al., 2019; Lorenz, 2016). My work identifies potential problems and sheds light on design opportunities for similar online spaces where interactions are real-time and ephemeral.

It is important to recognize that moderation is not one-size-fits-all. My findings point out that existing moderation strategies in one type of technology can break down completely in another. Therefore, while it may be easy and intuitive to import existing rules to a new community, designers and moderators should not ignore the technological infrastructure of the community when doing so, and carefully consider the limitations imposed by it.

Finally, it can be difficult to predict how people will abuse new technology, nor how rules or enforcement practices may need to change to prevent such abuse. Therefore, it is important that moderators are willing to change rules or make new rules for stakeholders with new needs, such as when communities adopt new technology. While my findings show that many of the rules were implicit in the context of new technology, I recommend moderators frequently reflect on their practices and consider whether implicit rules should be made explicit, so that new and old members can easily learn the rules.

6.2 Moderating Global Users

As social platforms operate at a global scale, so does their moderation. Therefore, online content moderation inevitably involves diverse groups of people. My empirical investigation in Chapter 4 reveals the complex reality of social media platforms' moderation of global users.

On a methodological level, researchers and practitioners may find significant value in adopting the free-text numeric measurement method that I used within their own platforms or communities. By assigning unbounded values to individual types of abusive behavior, my method is able to reveal precise, quantified relationships between them (e.g., one case of child exploitation imagery is equivalent to x cases of commercial spam in terms of severity). While reductive and not able to capture the full nuance of abusive behavior, these equivalence relationships can be a promising first step in guiding decisions in large-scale content moderation, including moderator resource allocation, proactive detection, and response priority for different kinds of content, especially given the inherent limitation in human moderation capacity (Gillespie, 2018a). Researchers and practitioners can also conduct the same measurement survey with diverse groups of stakeholders, such as users, moderators, and policy experts, to gain a variety of perspectives. In the context of platform moderation, such a multi-stakeholder measurement may prove especially fruitful, because regular users who rarely encounter highly severe types of abuse may not be able to reasonably estimate them.

On a theoretical level, my work contributes to a growing body of research against a one-size-fits-all approach to platform content moderation. As tempting as it is to devise an approach that "transcends" differences, the complexity of how people perceive abusive online behavior across the world indicates that it is unlikely for one approach to be desirable for everyone. As Schonebeck et al.

(2020) speculated, "it is likely that a monolithic approach to governance further magnifies inequities when applied in global, cross-cultural contexts." My findings point to several implications for how we can be better at engaging with these differences.

The finding that global users have differing perceptions of abusive behavior resonates in many ways with Irani et al.'s (2010) analytical framework of postcolonial computing that focuses on "reconfigur[ing] design-oriented cultural encounters." Through an example of a research making an unreasonable suggestion to Australian indigenous people due to a lack of cultural knowledge, Irani et al. voiced their concern of porting and translating knowledge while the infrastructure of knowing and telling may be different across cultures. Irani et al. noted "knowledge sharing—what it means to know something, and what it means to be able to tell it—is hemmed in all around by a series of infrastructures, social, cultural, and technological, that must be brought into alignment."

The issue of content moderation is exemplary of knowledge sharing and porting because it aims to impart often U.S.-created concepts and terms in community guidelines to other parts of the world. While social media platforms often use single pieces of community guidelines to govern their entire user bases, it is unclear whether the all-governing community guidelines written in English are capable of representing concepts in different languages, or whether the terms in English have equivalent translations to different languages. Therefore, designers and policymakers should consult with local experts and critically consider the local meanings of community guidelines when they are translated. As the social environment and the social media landscape constantly change, social media platforms may need to conduct longitudinal, ongoing field research to deeply understand whether the (translated) community guidelines are adequate for governing localized social media practices. It is critical for social media platforms to ensure that rules and intentions are properly communicated

across languages, and that users can participate in content moderation in a way that is not unintentionally harmful. Furthermore, as moderation has become increasingly distributed, moderators may have different personal, likely cultural understandings than the users whose content they are reviewing. Therefore, social media platforms should also consider improving existing ways to match moderators and content, as well as incorporating different local meanings into their moderator training to improve alignment of rule understandings.

The study in Chapter 4 also has important ethical implications. While I have identified several types of behavior that were regionally sensitive and important, I emphasize that there is no evidence for why they are important. I also intentionally avoided speculating possible reasons for these findings. Irani et al. (2010) argued for a generative, rather than taxonomic, view of culture that positions people at the intersection of cultures. Even though I conducted the study with some inherent regional taxonomies, geographically co-located individuals still experience and produce different, dynamic, and sometimes overlapping cultures (Irani et al., 2010). Therefore, reckless assumptions and speculations about "cultures" may not only be incorrect, but also highly dangerous, because they may impose harmful stereotypes and stigma on people. In the case of content moderation, the stake is even higher because the resulting stigma may be associated with highly severe, possibly illegal behavior. Therefore, I caution against simplistic interpretations of these regionally different topics, such as assigning cultural stereotypes to these differences (e.g., "Vietnam users perceive Creep Shots less severely than other regions because of their certain cultural values"). I urge researchers to conduct careful, deep-dive case studies into particular types of behavior to understand their full complexity.

Nevertheless, the study in Chapter 4 provides insights into the voice of international users in platform policymaking, which has been largely centered around compliance with various laws and regulations. While it is unclear how exactly moderation policies were made, or whether user inputs were taken into account in the making of these policies, given that platforms currently do not have customized policies for different geographical regions, it is reasonable to suspect that U.S. perspectives are weighed more heavily than those from other parts of the world. My research complicates and challenges the current practice of blanket policy making for the whole world by revealing the diverse attitudes toward abusive content from global users.

However, it is critical to note that the approach or the findings in Chapter 4 should not be used as the sole reason to determine whether or not a piece of content should be moderated. Nor the findings alone determine the degree of harshness to which one should be punished for posting abusive content. I raise these precautions for two reasons: First, the question that why people perceived the abusive behaviors as such remains unanswered. Second, more importantly, the perceptions of regular internet users should not be the gold standard of policymaking or the right way forward.

The complexity of analyzing abusive content globally may be exacerbated by the inability to describe and interpret regionally, reflecting Fricker's (2007) notion of hermeneutic injustice. Fricker defines hermeneutic injustice as when someone "has a significant area of their social experience obscured from understanding owing to prejudicial flaws in shared resources for social interpretation." In other words, a group of people may not have language for, and thereby not have the ability to describe and interpret a nuanced social experience that they do not share widely. Fricker uses the example of stalking to demonstrate the concept of hermeneutic injustice: A man

does not perceive stalking of a woman as harmful because the woman cannot describe her experience in the man's own interpretive system, and the man who never experienced stalking may have no idea what harm stalking actually entails. The same hermeneutic injustice is likely to apply in the context of content moderation as well: For example, people who are not aware of the consequences of anorexia nervosa may perceive its depiction as simply someone being "skinny" and therefore do not see it as being harmful.

Hermeneutic injustice also has broader global consequences, because much of the platform policymaking relies on translation rather than local development. While platform policies do exist in multiple languages, as opposed to being developed organically by people across the world who are familiar with local contexts, they are often translated from something developed by U.S. people speaking English, and essentially become different versions of U.S. ideals, not to mention how much might be lost in the translation. The survey in Chapter 4 suffers from the same epistemic limitation: It asked users to rate scenarios containing abusive behavior, but these scenarios, as well as the abusive behavior categories from which these scenarios arise, are conceptualized and described by people in a small part of the world (in this case, urban United States). People in the U.S. may perceive these categories and scenarios as comprehensive and representative, but it is possible that they do not even start to describe the harm and abuse that people in other parts of the world experience -- it may not even be possible to describe them in English. It is also likely that people in other parts of the world cannot understand the harm that the U.S.-centric policies describe because they do not exist in their interpretive systems. Therefore, while my findings reveal that global perceptions of harmful content can be different, they should not be the predominant factor in policymaking. Content moderation practices cannot fully protect people if they are only imparting some idealized version of the harms

these people experience, and policies for any geographical region should be developed locally in its own terms.

Beyond hermeneutic injustice, sometimes users' values may simply go against the values that a platform or the society holds, in which case it is in the platform's best interest to not take these user inputs into account in making platform policies. The possible value difference also raises broader questions for the user-centered tradition in HCI. While HCI has historically valued soliciting user inputs and fulfilling them, there is no guarantee that user perceptions are always going to be beneficial (or even legal), especially in determining what content is welcome or unwelcome on a platform. More likely than not, platforms will not want to fulfill users who want to see child pornography, who supports ethnic cleansing, or who thinks anything that does not pertain to a certain religious belief should be removed. While these are extreme examples, some of the surprising findings in Chapter 4, such as the depiction of eating disorders being lowly-rated in the United States, indicate that it is possible for some users to hold values that platforms may determine as harmful by value or by law. Furthermore, these values may become less apparent when they are aggregated into large amounts of user inputs, but still have the potential to deprioritize harms that the platform may want to be prioritized.

Therefore, I argue that while user inputs should be considered in policymaking, it is only part of a larger equation that will inevitably involve various balancing perspectives in legal, medical, political, and other localized domains. Content moderation is an example of when user-centered design hits its limits. Regular internet users are likely not well-equipped to provide input for high-danger, high-stake content moderation situations—their desires may be against the law, or cause harm in unintentional (or intentional) ways. While it may be valuable to understand user inputs, it is

more important to have actual experts with the requisite knowledge to keep these user inputs in check.

My call for attention to different people applies not only to commercial moderation of whole platforms, but to volunteer moderation of smaller communities as well—for example, moderating a gaming community should be different than moderating an LGBTQ community. Even with the same type of communities, rules and etiquette may still differ across individual communities (Dym & Fiesler, 2020). The challenges of moderating different people is an inevitable outcome as online communities have grown to the scale at which they operate today, and it is important to understand the community differences in norms and vulnerabilities in order to make moderation decisions.

6.3 Trade-offs Define Content Moderation

In Chapter 5, I have shown that a trade-off-centered framework significantly changes our view of content moderation: When we consider many content moderation cases side by side, existing moderation strategies may no longer work, agreed-upon rules may seem contentious, and the "right" ways forward may become in tension with other legitimate concerns.

Chapter 5 describes many competing choices in trade-offs in moderation actions and styles. Each choice has its own pros and cons that, as I have shown, relate to trade-offs in moderation philosophies. For example, the trade-off between leniency and harshness and that between immediately removing harm and long-term education in moderation actions demonstrate clear connections to the level of activity vs. quality of contribution and the nurturing vs. punishing trade-offs respectively. The different pros and cons of competing moderation styles also find their way to

trade-offs in philosophies. Overall, moderation philosophies reflect the fundamental needs and purposes that moderation actions and styles aim to serve.

In trade-offs in moderation philosophies, many options are often believed to (or at least be supposed to) go hand in hand with each other: Moderation should be both educational for sincere community members and punishing for malicious actors. Moderation should be both efficient and of high quality. Moderation should maintain community members' engagement and activities while ensuring a high quality of contribution. While these goals often seem to be congruent, in my analysis of moderation literature, I found that they were often at odds with each other. As ideal as it would be to achieve both sides of the trade-offs, I saw evidence that a tendency toward one side may necessarily be at the cost of the other.

Furthermore, these philosophies trace back to the very definition of content moderation. Grimmelmann (2015) defines moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." The trade-offs in moderation philosophies echo the goals of moderation in Grimmelmann's definition: Nurturing, moderation quality, and level of activity are different facets of facilitating cooperation, while punishing, moderation efficiency, and quality of contribution represent different dimensions of preventing abuse.

However, while Grimmelmann (2015) indicates that these two goals are to be achieved at the same time, my trade-off centered analysis shows a different relationship: facilitating cooperation and preventing abuse are at tension with each other in practice. If the two definitional components of content moderation constitute a trade-off, then I argue that content moderation as a whole can be

conceptualized as a series of trade-offs, and that at the core of moderation work is to make choices between simultaneously desirable goals.

Then, how can we balance facilitating cooperation and preventing abuse? The trade-offs in values may provide answers to this question. My work suggests that the driving force behind which component is favored more is dependent on the moderators' perceived identities of themselves and their visions for their communities, both of which are also shaped by various stakeholders including other moderators, community members, platforms, legal requirements, etc. (Gillespie, 2018a) These forces work together and converge toward a unique balance between facilitating cooperation and preventing abuse.

6.4 Acknowledge and Engage with Stakeholders and Trade-offs

The typical conceptualization of content moderation has largely been a jiu-jitsu between "moderators" and "users," but my dissertation has shown that the content moderation involves many more stakeholders—for example, human rights workers, journalists, etc. (Banchik, 2020), Even within "moderators" and "users," there are also different needs when we consider factors like demographics, gender identities, cultural backgrounds, and community infrastructures, and it would be ill-advised to assume that any one group can be treated uniformly. By examining multiple stakeholders side by side, my dissertation shows that a multi-stakeholder- and trade-off-centered perspective in content moderation is necessary and valuable, and taking these perspectives means that to acknowledge the needs of different stakeholders (and that they cannot be fully met at the same time), make conscious trade-offs between those needs, and ultimately be accountable to all stakeholders.

The world is not perfect—there are technologies that text-based moderation cannot handle; moderation resources are limited; some needs and desires have to be sacrificed. However, moderation needs to happen. As attractive as it is to be able to only identify and work toward the needs of specific stakeholders and situations that we resonate with, we cannot pretend that other stakeholders do not exist, or that their desires are consistent with those we already know about. It is bad to have problems occur for some stakeholders, but it is worse to not know or not acknowledge these problems even exist.

Discounting stakeholders can also happen unconsciously—it is easy for someone (with the best intention) to overlook stakeholders beyond their primary field of work or expertise. However, I argue that such unconscious overlooking of stakeholders is exceptionally dangerous, not only because the fact of overlooking itself might create potential for harm, but also because the assumption that all stakeholders are considered might result in blind confidence in decisions made from incomplete information. The former is obvious and remediable. The latter, however, takes deliberate, conscious efforts to overcome.

Given the importance of taking different stakeholders into account, the natural corollary is that content moderation should be fully personalized—each individual gets their own content moderation. However, this north star is built on the obviously unrealistic assumption that moderation decisions on an individual have no impact on other community members, not to mention that it is hardly scalable. Therefore, content moderation is inherently a multi-stakeholder one, and the very practice of content moderation may be in tension with the scale of content and people we want to support in online communities.

Then, trade-offs are inevitable when stakeholders have competing needs (which is nearly always the case), and the trade-off-centered framework in Chapter 5 offers a broad taxonomy of what the trade-offs are. If trade-offs go hand in hand with stakeholders, then no decision is a simple "decision" anymore. Every decision is part of a trade-off between stakeholders. Every decision toward something means deprioritizing and sacrificing something and someone else, but it is not always obvious what is being deprioritized, and whether that deprioritization is one that we can afford.

While it is not possible to fully customize content moderation, there is still significant value in customizing content moderation to some extent, given the abundance of evidence against a one-size-fits-all approach. While one may argue that customized moderation goes against the ideal of consistency, the currently common one-size-fits-all approach is the outcome of the pursuit of consistency, and it has failed on many fronts.

Therefore, while I am not indicating that we should abandon consistency as an evaluation criterion for content moderation, I argue that we should reconceptualize what consistency means. The predominant interpretation of consistency is that everyone should be governed by the same rules (possibly worded exactly the same), and enforced in exactly the same ways, and we are already seeing that such interpretation can be problematic because it ignores local contexts. Instead, we should reconceptualize consistency toward the goal of content moderation to reduce harm, and make it so that the degrees to which harm is reduced are similar across different stakeholders, which would require some level of customization. To what extent content moderation should be customized, however, depends on the specific stakeholders involved, resources available, among other factors. It is an open question that deserves further research.

While it is certainly impossible to come up with a perfect solution for prioritizing needs and customizing moderation (nor should we be trapped in trying to do so), I argue that a first step toward such a multi-stakeholder perspective is for moderators, designers, policymakers, and researchers to know what they are doing by asking themselves the following questions:

1. Who are the stakeholders?
2. Challenge their own assumptions by considering multiple dimensions along which the population can be categorized, and ask again: Who are the stakeholders?
3. What are the consequences of their decisions or propositions for the stakeholders? What are the costs and benefits?
4. Do their decisions or propositions privilege someone's perspective and disadvantage someone else's?
5. Is the result a privilege/disadvantage and cost/benefit ratio that they can reasonably accept?

While there is no lack of people who are good at coming up with ideas and making decisions, "knowing what they are doing" can take them one step further—it means to also be aware of the scope of the problem for which decisions are made, as well as the full consequences of these decisions that often propagate further than one would expect. Here I echo the points that Sultana et al. (2018) made through their research with women in rural Bangladesh, where they rightfully point out that blindly forcing the moral values that we take for granted without considering local situations could end up endangering the people we think we are helping—for example, designs that aim to destroy the deeply patriarchal society in Bangladesh could "empower" women into real social and physical harm. It is critical that we do not see stakeholders, whether they be Bangladeshi women

or people involved in moderation, as "objects of pity" that have no agency and are waiting to be saved—it is entirely possible that large-scale automated moderation will save commercial moderators from emotional trauma by driving them out of their job that they take pride in (Karunakaran & Ramakrishan, 2019) and actually pays a living wage.

I am not arguing whose perspective should be prioritized, but that prioritization itself should be prioritized. The questions above construct a systematic way to make informed moderation decisions. I have been arguing that we need to rethink the dimensions on which we categorize stakeholders, and just like how we as 3-dimensional beings cannot visualize 4-dimensional spaces, being constrained to the typical dimensions of stakeholders (e.g., moderators vs. users) prohibits decision makers from seeing the higher-dimensional space of stakeholders. However, by asking these questions, decision makers will be more likely to identify not only categories of stakeholders they previously overlooked, but also dimensions that make various categories possible.

After identifying the stakeholders, decision makers can then acknowledge rather than dismiss stakeholders' needs, embrace the hard problem of making trade-offs between their needs, and be cognizant of the non-obvious downstream consequences of their decisions. While the decisions will likely be imperfect, or sometimes even turn out to be harmful after a period of time, having the answers to these questions will help decision makers be accountable to and aware of their fundamental assumptions, and consequently iterate on and improve their decisions.

6.5 Future Research Directions

My dissertation is a first step in taking a multi-stakeholder perspective for understanding and improving content moderation, as well as demonstrating the importance of doing so. In this section,

I highlight some of the potential future research directions that are worth pursuing. I first explore opportunities for research in moderating more diverse kinds of emerging technologies on different platforms, and argue for anticipating potential harm to be a central component of system design. Then, I discuss additional ways for future research to understand harm and severity, as well as the need to expand content moderation research to a global scale. I close with a discussion of participatory design as a potential research paradigm for content moderation, and argue for a future of deep collaboration between academia and industry.

Moderating Emerging Technologies

First, my research in Chapter 3 suggested that recording and surveillance can be helpful in moderating real-time, ephemeral content, and platforms are indeed increasingly adopting this strategy. For example, Sony announced that PlayStation 5 will record voice chat audio for moderation purposes (McWhertor, 2020), extending the recording practice in a select few communities to their whole PlayStation platform. Facebook's new social VR platform, Horizon, will have every user recording other users on a rolling basis, thereby implementing both recording and "spying" (Lang, 2020).

While research on moderators suggests the benefits of such surveillance techniques, it will be worthwhile to study the impacts of these practices on another group of stakeholders: users. Do surveillance practices reduce harm in the community? Do they improve community members' behavior? What impacts do they have on the community norms? They are not only intuitive research questions to pursue, but taken together, they will also construct a broader, longitudinal picture of ephemeral, real-time content moderation.

On a higher level, Chapter 3 has also shown that the affordances of new technologies could present unforeseen challenges for content moderation that require non-trivial solutions. Therefore, rather than trying to fix them when breakdowns happen, a promising new design paradigm is to incorporate moderation as an essential factor for new technologies. The constant reports of abuse on new technologies suggest that it may be necessary to anticipate "the worst of humanity" in the design of them, and actively explore ways to prevent such abuse. Software engineering already has established procedures of testing for edge cases by intentionally trying to break the software (hence the joke, "a software testing engineer walks into a bar and orders -1 beers"), and I argue that "edge case testing" should also be a standard process in designing content moderation for socio-technical systems. Because abuse, by its nature, would never happen if everyone follows a system's intended "normal" use, active abuse prevention and content moderation will require designers to consider potential use cases in seemingly the most impossible scenarios.

While there are certainly many factors to consider in designing the corner cases for testing, I provide some initial questions here: Are there features that are easily exploitable (i.e., the "-1 beers") for known types of abuse? What about non-regular user-facing features (such as APIs)? Will unintended use of multiple features at the same time produce potential harm? Are there features that may not easily cause harm alone, but have potential for harm when used with third-party software? While new modalities of online communities (e.g., voice, live video, or VR) are straightfoward examples of new opportunities for harm by allowing new interactions, new points of exploitation of existing modalities (e.g., text) remain but may not be obvious, and they should deserve equal, if not more attention. While I realize abuse prevention measures from such corner case testing may

intervene with regular, intended use, each system is also different and unique, and researchers and designers of these systems should carefully consider where to draw the line on a case-by-case basis.

Further, as moderators in different communities use different strategies to moderate, it is likely that moderators in one community already have solutions to problems in another community. Therefore, it is worthwhile to test and evaluate a "moderator alliance" approach for interrogating and mitigating moderation problems. A possible approach is to form a community of moderators and schedule regular meetings where moderators share strategies and solutions to problems, which are also documented in a central knowledge repository. Research can then explore both the short-term and long-term effects of the moderator community (e.g., changes in moderator experience and workload in their respective communities). If the "moderator alliance" approach proves to be effective, it will not only benefit moderating emerging technologies but also facilitate moderating new problems in existing technologies.

The Complexity of the Severity of Harm

A main contribution of the research in Chapter 4 is a new way to measure the severity of harm. However, it is also a reductive approach that operationalizes severity in two dimensions, punishment and urgency, and only from the viewers' perspective—it asks people to rate abusive behavior when they see it on the platform as an outsider. It also constrains the harm on a single platform, Facebook. However, both harm and its severity are more complex than the two constructs and the one perspective I took in this research, and also happen on multiple platforms.

Therefore, a promising area of future research is to explore the full complexity of harm and severity. Ideally, future research will uncover multiple dimensions and perspectives that paint a full picture of harm and severity, which might include the scale of the harm, intent of the perpetrator,

and vulnerability of the victim. A full picture of harm will also further testify to the multi-stakeholder perspective of this dissertation by showing the same harm is experienced differently by different people.

With a fuller understanding of harm and severity, an intuitive next step would be to understand these broad concepts quantitatively by replicating the study in Chapter 4 while incorporating more dimensions and perspectives of harm. There are many ways to replicate the study with changes in key operationalizations, or "parameters" of the study, in dimensions, perspectives, and platforms. For example, one could incorporate new dimensions: Would the results change if other dimensions of severity (e.g., scale) are considered, in addition to punishment and urgency? How would results change across different dimensions of severity? One could also tweak the perspective from which the harm is viewed: Would the measured severity change if the participants are asked to take the perspective of the victim? Finally, one could also replicate the study on other platforms to see whether platform-specific effects exist: would results change if the abuse was posted on Nextdoor or livestreamed on YouTube?

These different options could result in a large number of combinations of parameters, so it will also be worthwhile to explore how one can meaningfully combine or collapse different parameters into evaluation metrics, which will be important for informing actual decision making in content moderation. As I mentioned in Section 6.2, user inputs alone cannot determine content moderation decisions, so in evaluating the efficacy of potential metrics, it will be important to involve other stakeholders such as legal and policy experts to fully understand the consequences of using one set of metrics vs. another. As the content moderation landscape constantly changes, the

development and evaluation of metrics will necessarily be continuing and iterative, instead of a one-off process.

Content Moderation in a Global Context

Chapter 4 also shows that moderation research and practices often have broad, international implications, but the design of many research studies and systems are often U.S.- or western-based (Barwulor et al., 2020). The variance in how international users perceive platform rules means that there need to be more international studies or studies that focus on other geographical areas, as online platforms are inevitably global. The fact that the majority of moderation research currently happens in Western contexts indicates a lack of awareness of the importance of content moderation in other areas. It will be valuable to replicate existing research in other regional and cultural contexts to discover any similarities or differences in content moderation practices and challenges. As Figure 5-1 shows, the surge of content moderation research is a fairly recent phenomenon, so it is reasonable to replicate most, if not all, exploratory moderation studies in other parts of the world, and proceed to further, more in-depth studies based on the specific research findings. For example, replicating Seering et al.'s (2019) work is likely to be informative because it will paint a broad picture of how community moderation functions in other parts of the world. On the other hand, replicating Jhaver et al.'s (2019) experimental work on moderation transparency would be too early, as there is currently not enough local context (which the former study would provide) to situate the highly-specific findings. At the same time, educating the whole international social computing community about content moderation will be a necessary line of future work to facilitate such global moderation research (e.g., holding workshops that intentionally reach out to international researchers).

Toward a Multi-stakeholder Future of Content Moderation

In the previous section, I called for an approach that consciously incorporates diverse stakeholders in the research process. The consideration of diverse stakeholders matters not only in conducting the research (e.g., collecting data from multiple stakeholders), but also in the analysis and recommendations provided, and it is easy to lose sight of the stakeholders in the latter part. Therefore, on a high level, I suggest that the research community widely adopt a practice of taking the recommendations that researchers think are good ideas back to the stakeholders and test their true effectiveness, and also to stakeholders beyond those with whom the original study was conducted to test how these recommendations work in other contexts. Researchers can carry out these tests at the scale of whole communities using infrastructures like CivilServant (Matias & Mou, 2018).

My suggestion above shares some similarities with participatory design in terms of (at least partly) involving stakeholders in the design process, but it remains unclear whether participatory design itself is an effective approach in content moderation research. One prominent concern is the limitation of users that I mentioned in Section 6.2. How productive will it be to put users and various experts in the same room to inform the design of various facets of content moderation? Related, would it be better to more heavily involve experts in the design of policies, but more heavily users in the design of user-facing tools? While the scale of social media platforms makes it impossible to involve all stakeholders, researchers may still involve stakeholders that cover a key range of issues and perspectives (e.g., user experience, moderation work, public safety, social justice, etc.) to guide exploring participatory design-inspired approaches as a new research paradigm in content moderation, and to evaluate its strengths and weaknesses.

It is important to note that many considerations in content moderation require specific kinds of familiarity and expertise that academic social computing researchers may not have—for example, legal constraints, labor relations, capacity management, etc., and each of these dimensions will introduce multiple categories of stakeholders. Therefore, I envision that the future of content moderation research necessarily needs to be an interdisciplinary one where academic researchers collaborate with industry practitioners. Collaborations with the industry will shed light on the inner-working of commercial moderation, which has been largely a black-box in academic research, and allow academic researchers to conduct empirical research that has been notably missing in the moderation literature. Further, such collaborations will also expose researchers in both academia and industry to previously overlooked stakeholders and their needs, and thereby encourage reflexivity about and consideration of these new stakeholders in their work.

6.6 Concluding Thoughts

I began my dissertation research to understand multi-stakeholder perspectives in online content moderation. Through an qualitative study of moderation in voice-based online communities on Discord and a quantitative study of global perspectives of abusive behavior, I have found that multi-stakeholders perspectives are varied and complex, and their competing interests are not easily reconcilable, but ignoring them can have serious consequences. By taking a step back and examining content moderation research more broadly through a systematic literature review, I have revealed that trade-offs are pervasive in how moderation is practiced in the real world, to the extent that it becomes core to the very definition of content moderation.

If multi-stakeholder trade-offs are inevitable in content moderation, then it is in our best interest to embrace them rather than to avoid them. Content moderation is a difficult problem

characterized by what Gillespie (2018a) calls "unresolvable tensions." It is easy for us to devote all efforts helping the stakeholders immediately visible in our mind and be satisfied with ourselves, but it often has negative or even harmful consequences to the stakeholders that we ignore in the process, and it might be too late to fix the damage by the time we realize their needs. If nothing else, I hope that my dissertation will serve as a first step for us to stop being complacent of ourselves in how we design and practice content moderation. Only by acknowledging various stakeholders and engaging with them can we develop responsible, actionable, and well-informed online content moderation.

references

Ackerman, M. S., Starr, B., Hindus, D., & Mainwaring, S. D. (1997). Hanging on the 'Wire: A Field Study of an Audio-only Media Space. *ACM Trans. Comput.-Hum. Interact.*, *4*(1), 39–66. https://doi.org/10.1145/244754.244756

Banchik, A. V. (2020). Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content. *New Media & Society*, 1461444820912724. https://doi.org/10.1177/1461444820912724

Barwulor, C., McDonald, A., Hargittai, E., & Redmiles, E. M. (2020). *"Disadvantaged in the American-dominated internet": Sex, Work, and Technology*. https://doi.org/10.31235/osf.io/vzehu

Bell, B. (2018, January 11). *VRChat Inc. Addresses Player Harassment Amid Surge in Player Base*. Pastemagazine.Com. https://www.pastemagazine.com/articles/2018/01/vrchat-inc-addresses-player-harassment-amid-surge.html

Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018, June 15). When Online Harassment Is Perceived as Justified. *Twelfth International AAAI Conference on Web and Social Media*. Twelfth International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17902

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.*, *1*(CSCW), 24:1–24:19. https://doi.org/10.1145/3134659

Blackwell, L., Ellison, N., Elliott-Deflo, N., & Schwartz, R. (2019). Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 100:1–100:25. https://doi.org/10.1145/3359202

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Caplan, R. (2018). *Content or context moderation? Artisanal, community-reliant, and industrial approaches* [White Paper]. Data & Society. https://datasociety.net/output/content-or-context-moderation/

Centivany, A., & Glushko, B. (2016). "Popcorn Tastes Good": Participatory Policymaking and Reddit's. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1126–1137. https://doi.org/10.1145/2858036.2858516

Chancellor, S., Baumer, E. P. S., & De Choudhury, M. (2019). Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 147:1–147:32. https://doi.org/10.1145/3359249

Chancellor, S., Hu, A., & De Choudhury, M. (2018). Norms Matter: Contrasting Social Support Around Behavior Change in Online Weight Loss Communities. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174240

Chancellor, S., Kalantidis, Y., Pater, J. A., De Choudhury, M., & Shamma, D. A. (2017). *Multimodal Classification of Moderated Online Pro-Eating Disorder Content*. 3213–3226. https://doi.org/10.1145/3025453.3025985

Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1171–1184. https://doi.org/10.1145/2818048.2819973

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). #Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1201–1213. https://doi.org/10.1145/2818048.2819963

Chandrasekharan, E., Gandhi, C., Mustelier, M. W., & Gilbert, E. (2019). Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), Article 174. https://doi.org/10.1145/3359276

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.*, *1*(CSCW), 31:1–31:22. https://doi.org/10.1145/3134666

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The Internet's Hidden Rules: An Empirical Study of Reddit Norm

Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), Article 32. https://doi.org/10.1145/3274301

Chen, B. X. (2018, August 7). Autoplay Videos Are Not Going Away. Here's How to Fight Them. *The New York Times*. https://www.nytimes.com/2018/08/01/technology/personaltech/autoplay-video-fight-them.html

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1217–1230. https://doi.org/10.1145/2998181.2998213

Child Pornography. (2015, May 26). https://www.justice.gov/criminal-ceos/child-pornography

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 24, pp. 201–234). Academic Press. https://doi.org/10.1016/S0065-2601(08)60330-5

Cibelli, K. (2017). *The Effects of Respondent Commitment and Feedback on Response Quality in Online Surveys* [University of Michigan]. https://deepblue.lib.umich.edu/handle/2027.42/136981

Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication. In L. Resnick, L. B, M. John, S. Teasley, & D. (Eds.), *Perspectives on Socially Shared Cognition* (pp. 13–1991). American Psychological Association.

Clement, J. (2019a, November 20). *Facebook users by country 2019*. Statista. https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/

Clement, J. (2019b, November 20). *Most used social media platform*. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Conger, K. (2019, October 30). Twitter Will Ban All Political Ads, C.E.O. Jack Dorsey Says. *The New York Times*. https://www.nytimes.com/2019/10/30/technology/twitter-political-ads-ban.html

Cyber Civil Rights Initiative. (n.d.). *46 States + DC + One Territory NOW have Revenge Porn Laws | Cyber Civil Rights Initiative*. Retrieved June 23, 2020, from https://www.cybercivilrights.org/revenge-porn-laws/

Datta, S., & Adar, E. (2019). Extracting Inter-Community Conflicts in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 146–157.

Diakopoulos, N., & Naaman, M. (2011). *Towards quality discourse in online news comments*. https://doi.org/10.1145/1958824.1958844

Dibbell, J. (1998). *My Tiny Life: Crime and Passion in a Virtual World*. Julian Dibbell.

DiSalvo, C., Sengers, P., & Brynjarsdóttir, H. (2010). Mapping the Landscape of Sustainable HCI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1975–1984. https://doi.org/10.1145/1753326.1753625

Discord. (2019). *Community Guidelines | Discord*. https://discord.com/guidelines

Donaldson, T., & Preston, L. E. (1995). The Stakeholder Theory of the Corporation: Concepts, Evidence, and Implications. *The Academy of Management Review*, *20*(1), 65–91. JSTOR. https://doi.org/10.2307/258887

Dosono, B., & Semaan, B. (2019). Moderation Practices As Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 142:1–142:13. https://doi.org/10.1145/3290605.3300372

Draper, N. A. (2019). Distributed intervention: Networked content moderation in anonymous mobile spaces. *Feminist Media Studies*, *19*(5), 667–683. https://doi.org/10.1080/14680777.2018.1458746

Duggan, M. (2017, July 11). Online Harassment 2017. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

Duguay, S., Burgess, J., & Suzor, N. (2018). Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine: *Convergence*. https://doi.org/10.1177/1354856518781530

Dym, B., & Fiesler, C. (2020). Ethical and privacy considerations for research using online fandom data. *Transformative Works and Cultures*, *33*. https://doi.org/10.3983/twc.2020.1733

Ehrett, J. S. (2016). E-judiciaries: A model for community policing in cyberspace. *Information & Communications Technology Law*, *25*(3), 272–291. https://doi.org/10.1080/13600834.2016.1236428

Einwiller, S. A., & Kim, S. (2020). How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation. *Policy & Internet*, *12*(2), 184–206. https://doi.org/10.1002/poi3.239

Facebook. (n.d.). *Community Standards*. Community Standards. Retrieved September 21, 2020, from https://www.facebook.com/communitystandards/

Facebook. (2019, December 20). Removing Coordinated Inauthentic Behavior From Georgia, Vietnam and the US. *About Facebook*. https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/

Fan, J., & Zhang, A. X. (2020). Digital Juries: A Civics-Oriented Approach to Platform Governance. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376293

Feuston, J. L., Taylor, A. S., & Piper, A. M. (2020). Conformity of Eating Disorders through Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 040:1–040:28. https://doi.org/10.1145/3392845

Fick, M., & Dave, P. (2019, April 23). Facebook's flood of languages leave it struggling to monitor content. *Reuters*. https://www.reuters.com/article/us-facebook-languages-insight-idUSKCN1RZ0DW

Fiedrich, F., Gehbauer, F., & Rickers, U. (2000). Optimized resource allocation for emergency response after earthquake disasters. *Safety Science*, *35*(1), 41–57. https://doi.org/10.1016/S0925-7535(00)00021-7

Fiesler, C., & Bruckman, A. S. (2019). Creativity, Copyright, and Close-Knit Communities: A Case Study of Social Norm Formation and Enforcement. *Proc. ACM Hum.-Comput. Interact.*, *3*(GROUP), Article 241. https://doi.org/10.1145/3361122

Fiesler, C., Jiang, J. "Aaron," McCann, J., Frye, K., & Brubaker, J. R. (2018, June 15). Reddit Rules! Characterizing an Ecosystem of Governance. *Twelfth International AAAI Conference on Web and Social Media*. Twelfth International AAAI Conference on Web and Social Media. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17898

Fiesler, C., Lampe, C., & Bruckman, A. S. (2016). Reality and Perception of Copyright Terms of Service for Online Content Creation. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, 1450–1461.

Freeman, R. E. (1983). *Strategic Management: A Stakeholder Approach*. Cambridge University Press. https://doi.org/10.1017/CBO9781139192675

Frey, S., Krafft, P. M., & Keegan, B. C. (2019). "This Place Does What It Was Built For": Designing Digital Institutions for Participatory Change. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 32:1–32:31. https://doi.org/10.1145/3359134

Fricker, M. (2007). Hermeneutical Injustice. In *Epistemic Injustice*. Oxford University Press. https://oxford-universitypressscholarship-com.colorado.idm.oclc.org/view/10.1093/acprof:oso/9780198237907.001.0001/acprof-9780198237907-chapter-8

Gallagher, S. E., & Savage, T. (2016). Comparing learner community behavior in multiple presentations of a Massive Open Online Course. *Journal of Computing in Higher Education*, *28*(3), 358–369. https://doi.org/10.1007/s12528-016-9124-y

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Getto, G., & Labriola, J. T. (2016). iFixit Myself: User-Generated Content Strategy in "The Free Repair Guide for Everything." *IEEE Transactions on Professional Communication*, *59*(1), 37–55. https://doi.org/10.1109/TPC.2016.2527259

Gibson, A. (2019). Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces: *Social Media + Society*. https://doi.org/10.1177/2056305119832588

Gilbert, S. A. (2020). "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 019:1–019:27. https://doi.org/10.1145/3392822

Gillespie, T. (2010). The politics of 'platforms.' *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2018a). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Gillespie, T. (2018b). Regulation of and by Platforms. In *The SAGE Handbook of Social Media* (pp. 254–278). SAGE Publications Ltd. https://doi.org/10.4135/9781473984066

Gillespie, T. (2019, October 2). Content Moderation: A Reading List. *Social Media Collective*. https://socialmediacollective.org/reading-lists/content-moderation-reading-list/

Gleckman, H. (2018). *Multistakeholder Governance and Democracy: A Global Challenge*. Routledge. https://doi.org/10.4324/9781315144740

Gray, J. E., & Suzor, N. P. (2020). Playing with machines: Using machine learning to understand automated copyright enforcement at scale. *Big Data & Society*, *7*(1), 2053951720919963. https://doi.org/10.1177/2053951720919963

Grimmelmann, J. (2015). The Virtues of Moderation. *Yale Journal of Law and Technology*, *17*(1). https://digitalcommons.law.yale.edu/yjolt/vol17/iss1/2

Grover, T., & Mark, G. (2019). Detecting Potential Warning Behaviors of Ideological Radicalization in an Alt-Right Subreddit. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 193–204.

Guilford, J. P. (1954). *Psychometric methods, 2nd ed*. McGraw-Hill.

Gurzick, D., White, K. F., Lutters, W. G., & Boot, L. (2009). A view from Mount Olympus: The impact of activity tracking tools on the character and practice of moderation. *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, 361–370. https://doi.org/10.1145/1531674.1531727

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. JSTOR. https://doi.org/10.2307/2346830

Harwell, D. (2018, July 19). *The accent gap: How Amazon's and Google's smart speakers leave certain voices behind*. Washington Post. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/

Heinze, A., Ferneley, E., & Child, P. (2013). Ideal Participants in Online Market Research: Lessons from Closed Communities: *International Journal of Market Research*. https://doi.org/10.2501/IJMR-2013-066

Holmes, P., & Cox, A. M. (2011). "Every group carries the flavour of the admins": Leadership on Flickr. *International Journal of Web Based Communities*, *7*(3), 376–391. https://doi.org/10.1504/IJWBC.2011.041205

Hossain, M. S. (2015). Social Media and Terrorism: Threats and Challenges to the Modern Era. *South Asian Survey*, *22*(2), 136–155. https://doi.org/10.1177/0971523117753280

Hua, Y., Naaman, M., & Ristenpart, T. (2020). Characterizing Twitter Users Who Engage in Adversarial Interactions against Political Candidates. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376548

Hudson, J. M., & Bruckman, A. (2004). "Go Away": Participant Objections to Being Studied and the Ethics of Chatroom Research. *The Information Society*, *20*(2), 127–139. https://doi.org/10.1080/01972240490423030

The Public Order Act 1986, 1986 c.64 (1986). http://www.legislation.gov.uk/ukpga/1986/64/contents

Internet Governance Forum. (2020, February 27). *IGF 2020 Thematic Tracks* [Text]. Internet Governance Forum. https://www.intgovforum.org/multilingual/content/igf-2020-thematic-tracks

Irani, L., Vertesi, J., Dourish, P., Philip, K., & Grinter, R. E. (2010). Postcolonial computing: A lens on design and development. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1311–1320. https://doi.org/10.1145/1753326.1753522

Isaacs, E. A., & Tang, J. C. (1994). What video can and cannot do for collaboration: A case study. *Multimedia Systems*, *2*(2), 63–73. https://doi.org/10.1007/BF01274181

Jackman, T. (2018, April 11). *Trump signs 'FOSTA' bill targeting online sex trafficking, enables states and victims to pursue websites*. Washington Post. https://www.washingtonpost.com/news/true-crime/wp/2018/04/11/trump-signs-fosta-bill-targeting-online-sex-trafficking-enables-states-and-victims-to-pursue-websites/

Jackson, S. J., Gillespie, T., & Payette, S. (2014). The Policy Knot: Re-integrating Policy, Practice and Design in Cscw Studies of Social Computing. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 588–602. https://doi.org/10.1145/2531602.2531674

Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019). "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), Article 192. https://doi.org/10.1145/3359294

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.*, *26*(5), Article 31. https://doi.org/10.1145/3338243

Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), Article 150. https://doi.org/10.1145/3359252

Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Trans. Comput.-Hum. Interact.*, *25*(2), 12:1–12:33. https://doi.org/10.1145/3185593

Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2018). "The Perfect One": Understanding Communication Practices and Challenges with Animated GIFs. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), Article 80. https://doi.org/10.1145/3274349

Jiang, J. A., Kiene, C., Middler, S., Brubaker, J. R., & Fiesler, C. (2019). Moderation Challenges in Voice-based Online Communities on Discord. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 55:1–55:23. https://doi.org/10.1145/3359157

Jiang, J. A., Middler, S., Brubaker, J. R., & Fiesler, C. (2020, October). Characterizing Community Guidelines on Social Media Platforms. *CSCW 2020 Companion*.

Jiang, S., Robertson, R. E., & Wilson, C. (2019). Bias Misperceived:The Role of Partisanship and Misinformation in YouTube Comment Moderation. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 278–289.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/b98835

Juneja, P., Rama Subramanian, D., & Mitra, T. (2020). Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction*, *4*(GROUP), 17:1–17:35. https://doi.org/10.1145/3375197

Juneström, A. (2019). Online user misconduct and an evolving infrastructure of practices: A practice-based study of information infrastructure and social practices. *Information Research: An International Electronic Journal*, *24*(1). http://informationr.net/ir/24-1/isic2018/isic1825.html

Karunakaran, S., & Ramakrishan, R. (2019). Testing Stylistic Interventions to Reduce Emotional Impact of Content Moderation Workers. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*(1), 50–58.

Kayhan, V. O., & Bhattacherjee, A. (2013). Content Use from Websites: Effects of Governance Mechanisms. *Journal of Computer Information Systems*, *53*(4), 68–80. https://doi.org/10.1080/08874417.2013.11645652

Keegan, B. C., & Fiesler, C. (2017). The Evolution and Consequences of Peer Producing Wikipedia ' s Rules. *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*.

Kiene, C., Jiang, J. A., & Hill, B. M. (2019). Technological Frames and User Innovation: Exploring Technological Change in Community Moderation Teams. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 44:1–44:23. https://doi.org/10.1145/3359146

Kiene, C., Monroy-Hernández, A., & Hill, B. M. (2016). Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1152–1156. https://doi.org/10.1145/2858036.2858356

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review*, *131*, 1598–1670.

Kou, Y., & Nardi, B. (2013). *Regulating anti-social behavior on the Internet: The example of League of Legends*. https://doi.org/10.9776/13289

Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., Konstan, J., Ren, Y., & Riedl, J. (2011). *Building Successful Online Communities: Evidence-Based Social Design*. Mit Press; JSTOR. https://www.jstor.org/stable/j.ctt5hhgvw

Kriegel, H.-P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, *3*(1), 1:1–1:58. https://doi.org/10.1145/1497577.1497578

Lampe, C., & Johnston, E. (2005). Follow the (Slash) Dot: Effects of Feedback on New Members in an Online Community. *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, 11–20. https://doi.org/10.1145/1099203.1099206

Lampe, C., & Resnick, P. (2004). Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 543–550. https://doi.org/10.1145/985692.985761

Lampe, C., Zube, P., Lee, J., Park, C. H., & Johnston, E. (2014). Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*, *31*(2), 317–326. https://doi.org/10.1016/j.giq.2013.11.005

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lang, B. (2020, August 28). In "Horizon" Facebook Can Invisibly Observe Users in Real-time to Spot Rule Violations. *Road to VR*. https://www.roadtovr.com/facebook-horizon-privacy-monitoring-moderation/

Lessig, L. (2006). *Code: And other Laws of Cyberspace, Version 2.0*. Basic Books.

Liao, Q., Pan, Y., Zhou, M. X., & Ma, F. (2010). Chinese online communities: Balancing managementcontrol and individual autonomy. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2193–2202. https://doi.org/10.1145/1753326.1753658

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine*, *6*(7). https://doi.org/10.1371/journal.pmed.1000100

Lorenz, T. (2016, May 26). *Here's What Happened When I Was Surrounded by Men in Virtual Reality*. https://mic.com/articles/144470/sexual-harassment-in-virtual-reality

Luo, M., Hsu, T. W., Park, J. S., & Hancock, J. T. (2020). Emotional Amplification During Live-Streaming: Evidence from Comments During and After News Events. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 047:1–047:19. https://doi.org/10.1145/3392853

Matias, J. N. (2016a). Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1138–1151. https://doi.org/10.1145/2858036.2858391

Matias, J. N. (2016b). The civic labor of online moderators. *The Platform Society*, 10.

Matias, J. N. (2019). The Civic Labor of Volunteer Moderators Online: *Social Media + Society*. https://doi.org/10.1177/2056305119836778

Matias, J. N., & Mou, M. (2018). CivilServant: Community-Led Experiments in Platform Governance. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 9:1–9:13. https://doi.org/10.1145/3173574.3173583

McWhertor, M. (2020, October 15). *PS5 can record voice chat audio for moderation purposes*. Polygon. https://www.polygon.com/2020/10/15/21517451/ps4-ps5-record-party-chat-audio-moderation-firmware

Medeiros, B. (2019). Picketing the Virtual Storefront: Content Moderation and Political Criticism of Businesses on Yelp. *International Journal of Communication*, *13*(0), 17.

Mitchell, R. C., Carson, R. T., & Carson, R. T. (1989). *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Resources for the Future.

Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts. *The Academy of Management Review*, *22*(4), 853–886. JSTOR. https://doi.org/10.2307/259247

Newell, E., Jurgens, D., Saleem, H. M., Vala, H., Sassine, J., Armstrong, C., & Ruths, D. (2016, March 31). User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. *Tenth International AAAI Conference on Web and Social Media*. Tenth International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13137

Newton, C. (2019, February 25). *The secret lives of Facebook moderators in America*. The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Nurik, C. (2019). "Men Are Scum": Self-Regulation, Hate Speech, and Gender-Based Censorship on Facebook. *International Journal of Communication*, *13*(0), 21.

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, *23*(1), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

Park, D., Sachar, S., Diakopoulos, N., & Elmqvist, N. (2016). Supporting Comment Moderators in Identifying High Quality Online News Comments. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1114–1125. https://doi.org/10.1145/2858036.2858389

Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. *Proceedings of the 19th International Conference on Supporting Group Work*, 369–374. https://doi.org/10.1145/2957276.2957297

Paul, S. (2017, March 20). Voice Is the Next Big Platform, Unless You Have an Accent | Backchannel. *Wired*. https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/

Pavalanathan, U., Han, X., & Eisenstein, J. (2018). Mind Your POV: Convergence of Articles and Editors Towards Wikipedia's Neutrality Norm. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 137:1–137:23. https://doi.org/10.1145/3274406

Pellicone, A. J., & Ahn, J. (2017). The Game of Performing Play: Understanding Streaming as Cultural Production. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4863–4874. https://doi.org/10.1145/3025453.3025854

Petrič, G., & Petrovčič, A. (2014). Elements of the management of norms and their effects on the sense of virtual community. *Online Information Review*, *38*(3), 436–454. https://doi.org/10.1108/OIR-04-2013-0083

Phadke, S., & Mitra, T. (2020). Many Faced Hate: A Cross Platform Study of Content Framing and Information Sharing by Online Hate Groups. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376456

Potts, L., Small, R., & Trice, M. (2019). Boycotting the Knowledge Makers: How Reddit Demonstrates the Rise of Media Blacklists and Source Rejection in Online Communities.

IEEE Transactions on Professional Communication, *62*(4), 351–363. https://doi.org/10.1109/TPC.2019.2946942

Procházka, O. (2019). Making Sense of Facebook's Content Moderation: A Posthumanist Perspective on Communicative Competence and Internet Memes. *Signs and Society*, *7*(3), 362–397. https://doi.org/10.1086/704763

Quarantined Subreddits. (n.d.). Reddit Help. Retrieved September 17, 2019, from https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/quarantined-subreddits

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 557–568.

Redmiles, E. M., Bodford, J., & Blackwell, L. (2019). "I Just Want to Feel Safe": A Diary Study of Safety Perceptions on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 405–416.

Riedl, M. J., Masullo, G. M., & Whipple, K. N. (2020). The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior*, *107*, 106262. https://doi.org/10.1016/j.chb.2020.106262

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Robertson, A. (2019, June 26). *Reddit quarantines Trump subreddit r/The_Donald for violent comments*. The Verge. https://www.theverge.com/2019/6/26/18759967/reddit-quarantines-the-donald-trump-subreddit-misbehavior-violence-police-oregon

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruckenstein, M., & Turunen, L. L. M. (2020). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society*, *22*(6), 1026–1042. https://doi.org/10.1177/1461444819875990

Saldaña, J. (2009). *The Coding Manual for Qualitative Researchers*. SAGE Publications, Incorporated.

Sarkar, C., Wohn, D., Lampe, C., & DeMaagd, K. (2012). A quantitative explanation of governance in an online peer-production community. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2939–2942. https://doi.org/10.1145/2207676.2208701

Schlesinger, A., Chandrasekharan, E., Masden, C. A., Bruckman, A. S., Edwards, W. K., & Grinter, R. E. (2017). Situated anonymity: Impacts of anonymity, ephemerality, and hyper-locality on social media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6912–6924. https://doi.org/10.1145/3025453.3025682

Schoenebeck, S., Haimson, O. L., & Nakamura, L. (2020). Drawing from justice theories to support targets of online harassment. *New Media & Society*, 1461444820913122. https://doi.org/10.1177/1461444820913122

Scott, M., & Isaac, M. (2016, September 9). Facebook Restores Iconic Vietnam War Photo It Censored for Nudity. *The New York Times*. https://www.nytimes.com/2016/09/10/technology/facebook-vietnam-war-photo-nudity.html

Seering, J., Kraut, R., & Dabbish, L. (2017). Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 111–125. https://doi.org/10.1145/2998181.2998277

Seering, J., Ng, F., Yao, Z., & Kaufman, G. (2018). Applications of Social Identity Theory to Research and Design in Computer-Supported Cooperative Work. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), 201:1–201:34. https://doi.org/10.1145/3274771

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, 1461444818821316. https://doi.org/10.1177/1461444818821316

Shen, Q., & Rose, C. (2019). The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy. *Proceedings of the Third Workshop on Abusive Language Online*, 58–69. https://doi.org/10.18653/v1/W19-3507

Shieber, J. (2020, May 29). Zuckerberg explains why Facebook won't take action on Trump's recent posts. *TechCrunch*. https://social.techcrunch.com/2020/05/29/zuckerberg-explains-why-facebook-wont-take-action-on-trumps-recent-posts/

Skousen, T., Safadi, H., Young, C., Karahanna, E., Safadi, S., & Chebib, F. (2020). Successful Moderation in Online Patient Communities: Inductive Case Study. *Journal of Medical Internet Research*, *22*(3), e15983. https://doi.org/10.2196/15983

Smith, H. W. (1980). A modest test of cross-cultural differences in sexual modesty, embarrassment and self-disclosure. *Qualitative Sociology*, *3*(3), 223–241. https://doi.org/10.1007/BF00987137

Squirrell, T. (2019). Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society*, *21*(9), 1910–1927. https://doi.org/10.1177/1461444819834317

Srinivasan, K. B., Danescu-Niculescu-Mizil, C., Lee, L., & Tan, C. (2019). Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 163:1–163:21. https://doi.org/10.1145/3359265

Sultana, S., Guimbretière, F., Sengers, P., & Dell, N. (2018). Design Within a Patriarchal Society: Opportunities and Challenges in Designing for Rural Women in Bangladesh. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3173574.3174110

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication*, *13*(0), 18.

Telephone call recording laws. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Telephone_call_recording_laws&oldid=889445182

The Danish Penal Code, 126 (1930).

Tyler, T., Katsaros, M., Meares, T., & Venkatesh, S. (2019). Social media governance: Can social media companies motivate voluntary rule following behavior among their users? *Journal of Experimental Criminology*. https://doi.org/10.1007/s11292-019-09392-z

U.S. Department of Justice. (2015, May 26). *Citizen's Guide To U.S. Federal Child Exploitation And Obscenity Laws*. https://www.justice.gov/criminal-ceos/citizens-guide-us-federal-child-exploitation-and-obscenity-laws

Vanian, J. (2017, October 27). *This Is How Twitter Plans to Combat Non-Consensual Nudity*. Fortune. https://fortune.com/2017/10/27/nudity-revenge-porn-twitter/

Vashistha, A., Cutrell, E., Borriello, G., & Thies, W. (2015). Sangeet Swara: A Community-Moderated Voice Forum in Rural India. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 417–426. https://doi.org/10.1145/2702123.2702191

von Hirsch, A. (1992). Proportionality in the Philosophy of Punishment. *Crime and Justice*, *16*, 55–98.

Wang, G., Wang, B., Wang, T., Nika, A., Zheng, H., & Zhao, B. Y. (2014). Whispers in the dark: Analysis of an anonymous social network. *Proceedings of the 2014 Conference on Internet Measurement Conference*, 137–150. https://doi.org/10.1145/2663716.2663728

West, S. M. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Wijeratne, Y. (2020, July 23). Facebook, language and the difficulty of moderating hate speech. *Media@LSE*. https://blogs.lse.ac.uk/medialse/2020/07/23/facebook-language-and-the-difficulty-of-moderating-hate-speech/

Witt, A. E. A., Suzor, N., & Huggins, A. (2019). The rule of law on Instagram: An evaluation of the moderation of images depicting women's bodies. *The University of New South Wales Law Journal*, *42*(2), 557–596.

Wohn, D. Y. (2019). Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300390

Yang, D., Kraut, R. E., Smith, T., Mayfield, E., & Jurafsky, D. (2019). Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300574

Zeng, J., Chan, C., & Fu, K. (2017). How Social Media Construct "Truth" Around Crisis Events: Weibo's Rumor Management Strategies After the 2015 Tianjin Blasts. *Policy & Internet*, *9*(3), 297–320. https://doi.org/10.1002/poi3.155

A
APPENDIX
semi-structured interview script for discord moderators

The following interview script was used in the study with Discord moderators in Chapter 3.

CONSENT

Who I Am, What I am Doing

My name is <name>, and I'm from the Department of Information Science at the University of Colorado Boulder. Thank you so much for spending the time and talk with me today. I would like to talk to you about your experience with moderation in [server name]. You should have already received a Study Information Sheet that contains detailed information about this study.

Privacy and Confidentiality

To clarify, your participation in this conversation is strictly confidential - what we talk about stays with us. You also absolutely can decline to answer any of the questions I ask. You are also free to stop the interview at any time, you just have to let me know.

Consent to Interview and Recording

For the record, is it okay if I record this conversation?

Participant Questions?

Do you have any questions about this interview before we start?

Explanation of The Open Interview Structure

Just want you to know that there are no wrong answers to my questions. I want to hear your stories and your experiences. This is your interview.

Demographics

1. Age
2. Gender
3. Country
4. How long have you been on Discord
5. How long have you been moderating <server>?

Interview Questions

1. What's your day-to-day life in moderating <server>?
2. If there is a specific thing to talk about: I noticed <this thing> in the <community>, could you explain why <this thing> happens/exists?
    a. If not: Can you tell me the most memorable event in your job as a moderator?
3. How many hours per day / per week do you spend moderating this <community>
    a. Alternative: What's your workload?
4. Could you please describe a typical day as a <community> moderator?
5. Were you around when this <community> was created?

    a. If so, could you describe how the <community> was formed?
    b. If Discord: How were channels decided on?
    c. How were the rules for the <community> formed?
    d. How did you use the platform's tools to organize the <community>?
    e. How were roles formed, and what do these roles mean?
    f. Did you use any tools not provided by the platform, like bots?
6. How would you describe your relationship with community members (non-moderator members)?
7. Have you ever had to change a rule or make a new rule?
    a. If yes, have the participant tell an example
8. Have you and other moderators ever disagreed on how to apply a rule?
    a. If yes, have the participant tell an example
    b. Could you describe how and if this disagreement was resolved?
    c. Have the rules ever changed, and if so, could you describe how and why this rule change took place?
    d. Could you describe other "behind-the-scenes" decision making and communication that may occur privately between moderators?
9. (If the <community> uses bots) How would you describe the role that your bot plays in this server?
    a. If yes, have the participant tell an example
    b. Are these bots produced from within the community or are they invited to the server externally?

    c. How are the features of the bot decided on?
    d. How much work goes into maintaining these bots?
10. (If Discord and voice channels are used) How are the voice channels for this server moderated?
    a. Could you tell me an example where you or another moderator had to regulate a voice channel?
11. How would you describe the relationship between the moderators and admins of this <community> with the <Platform> moderation staff?
    a. Have platform-wide rules ever had an impact on your community?
12. Is there anything else you would like to talk about or comment on?
13. Finally, we would like to thank you for participating in this study by sending you a $20 Amazon Gift Card. Is <participant's contact info> still the best way to contact you?

B

APPENDIX
community guideline content analysis

	Facebook	YouTube	Instagram	Tik Tok	Reddit	Twitter	LinkedIn	Snapchat	Pinterest	Viber	Discord
Adult Non-Consensual Intimate Imagery	•	•	•	•	•	•	•	•	•	•	•
Adult Non-Sexual Nudity	•	•	•	•		•	•				
Celebrating Own Crime	•						•				
Child Exploitation Imagery	•	•	•	•	•	•	•	•	•	•	•
Child Nudity	•	•	•	•	•		•	•	•	•	
Coordinating harm	•	•	•	•	•	•	•	•	•		
Creep Shots	•	•			•	•	•				
Criminal Group Coordination	•	•	•	•		•	•		•		
Criminal Group Propaganda	•	•	•	•		•	•		•		

Cruel and Insensitive	•	•								•			
Digital Nudity	•	•	•	•		•	•						
Distribution of Virus	•			•			•			•			
Eating Disorder Depiction	•	•	•	•		•	•			•	•	•	
Eating Disorder Promotion	•	•	•	•		•	•			•	•	•	
Engagement Abuse	•	•	•		•	•	•						
False News and Misinformation	•	•				•				•			
Fraud and Financial Scams	•	•		•			•				•	•	
Graphic Violence: Animal Abuse	•	•	•	•		•	•	•		•	•	•	
Graphic Violence: Child Abuse	•	•	•	•		•	•	•		•	•	•	
Graphic Violence: Mutilated Humans	•	•	•	•		•	•	•		•	•	•	
Harassment and Bullying	•	•	•	•	•	•	•	•		•	•	•	
Hate Group Coordination	•		•	•		•	•			•			
Hate Group Propaganda	•		•	•		•	•			•			
Hate Speech: Dehumanization	•	•	•	•		•	•	•		•	•		
Hate Speech: Exclusion/Segregation	•	•	•	•		•	•	•		•	•		
Hate Speech: Inferiority	•	•	•	•		•	•	•		•	•		
Hate Speech: Slurs	•	•	•	•		•	•	•		•	•		
Hate Speech: Violence	•	•	•	•	•	•	•	•		•	•		
High Profile Impersonation	•	•		•	•	•	•	•		•	•		
Human Trafficking	•	•	•			•							

Inappropriate Interactions with Children	•	•	•	•	•	•	•	•	•	•		
Inauthentic Behavior	•	•	•	•	•	•	•	•	•	•	•	
Intellectual Property Infringement	•	•	•	•		•			•	•	•	
Interrupting Platform Services				•	•				•			
Mass Murder Coordination	•	•	•		•	•	•		•			
Mass Murder Support	•	•	•		•	•	•		•			
Minors sexualization	•	•	•	•	•	•	•	•	•	•	•	
Non-Consensual Intimate Imagery Threat	•	•	•		•	•	•					
Non-Consensual Sexual Touching	•	•	•	•	•	•	•		•			
Privacy Violation	•	•	•	•	•	•			•	•		
Private Impersonation	•	•		•	•	•	•	•	•	•		
Prostitution	•	•	•	•		•	•		•			
Regulated Goods: Alchohol and Tobacco Sale	•		•		•				•			
Regulated Goods: Endangered Species Sale	•		•			•			•			
Regulated Goods: Firearm Sales	•	•	•	•	•	•		•	•			
Regulated Goods: Human Organ Sale	•											
Regulated Goods: Live Animal Sale	•		•									
Regulated Goods: Marijuana Sales	•	•	•	•	•	•		•	•			
Regulated Goods: Non-medical Drug Sale	•	•	•	•	•	•	•	•	•			
Regulated Goods: Non-medical Drug Use	•	•		•			•		•			

Regulated Goods: Pharmaceutical Sales	•	•	•	•	•	•	•	•	•			
Sadism/Glorifying Violence	•	•			•		•		•	•		
Self-injury Depiction	•	•	•	•		•	•	•	•	•	•	
Self-injury Promotion	•	•	•	•		•	•	•	•	•	•	
Sexual Activity	•	•	•	•		•	•	•	•	•		
Sexual Solicitation	•	•		•		•	•		•	•		
Sexually Explicit Language	•	•		•			•		•	•		
Spam	•	•	•		•	•	•	•	•		•	
Suicide Depiction	•	•	•	•		•	•		•	•	•	
Suicide Promotion	•	•	•	•		•	•		•	•	•	
Terrorism Coordination	•	•	•	•		•	•	•	•			
Terrorist Propaganda	•	•	•	•		•	•	•	•			
Theft	•	•	•				•					
Vandalism	•		•				•					
Violence and Incitement	•	•	•	•	•	•	•	•	•		•	
Voter Fraud and Suppression	•	•				•						

Table B-1. Results from the community guideline content analysis.

C
APPENDIX
severity survey instrument

The following is the English version of the survey instrument used in the Phase 2 of Chapter 4. The other language versions are not included in this dissertation.

Commitment Screener

We care about the quality of our data, in order for us to get the most accurate measures of your opinions, it is important that you thoughtfully provide your best answers to each question in this survey.

Do you commit to thoughtfully provide your best answers to each question in this survey?

- ○ I will provide my best answers
- ○ I will NOT provide my best answers
- ○ I can't promise either way

Social Media Use Screener

Which of the following social media platforms you have used **in the last 30 days**? Please select all that apply.

- ☐ Twitter

☐ Snapchat

☐ Reddit

☐ Facebook

☐ Instagram

☐ YouTube

☐ WhatsApp

☐ None of the above

Instruction

Welcome and thank you for agreeing to participate in this research study. Your participation is important to us, and we are thankful for your time and effort.

Please read carefully.

Throughout this survey, we will show you examples of different kinds of online content. Along with each example there will be three questions to answer about that content. The examples have been formatted to look like posts from one of social media platforms you have used in the last 30 days: **Facebook.**

For the first question, you will need to decide how much money you would fine the person who posted the content as a punishment. The amount of money fined should reflect the penalty you feel this person deserves for posting the content. **In other words, the worse you feel the content is, the more you should fine the person.**

For the second question, you will need to decide how much money Facebook should spend to make sure the content gets removed immediately over other types of content. The amount of money

should reflect the urgency of reviewing and potentially removing that content. In other words, **the worse you feel the content is, the more Facebook should spend to prioritize it.**

For the third question, you will need to answer how upsetting the content is to you personally.

For the first two questions, **there is no limit to the amount you may decide. Zero is also an acceptable answer** if you don't think the content is bad or violates any Facebook rules.

As a reminder, your participation is voluntary. Your responses are anonymous and you may choose to stop participating at any time.

Before we begin the main survey, we would like you to complete some practice examples. In the practice as well as the main survey, the content examples will vary, but the questions to answer will always be the same. The button below will take you to the first practice question.

Scenario Questions

For the practice questions and the main survey, participants will answer the following three questions for every scenario presented. Note that "{{local currency}}" will be replaced with the currency of the participant's country. For example, U.S.-based participants will see "U.S. Dollars (USD)," and Vietnam-based participants will see "đồng Việt Nam (VND)."

- How much money, if any, would you fine the person who posted this content? Please indicate your answer in {{local currency}}. You only need to enter a number. (Free-text response)
- How much money, if any, do you think Facebook should spend to remove this content immediately over other types of content? Please indicate your answer in {{local currency}}. You only need to enter a number. (Free-text response)

- How upsetting is this content to you, if at all?
    - Extremely upsetting
    - Very upsetting
    - Somewhat upsetting
    - A little upsetting
    - Not at all upsetting

Practice Scenarios

Every participant will see the following two practice scenarios, and answer the three questions in Scenario Questions for each scenario.

1. Imagine you saw: **a wedding photo** on Facebook
2. Imagine you saw: a post of **congratulations on a friend's job announcement** on Facebook.

For these two scenarios, if the participant gives a non-zero answer to the first question (fine), they will see the following text on the next screen:

You chose to fine the person in the last practice question. Did you think the content was bad or violated Facebook rules? Remember, **feel free to enter if you don't think the content is bad or violates Facebook rules.**

If the participant gives a non-zero answer to the second question (Facebook spending), they will see the following text on the next screen:

You indicated that Facebook should spend money to remove the content immediately in the last practice question. Did you think the content was bad or violated Facebook rules? Remember, **feel free to enter if you don't think the content is bad or violates Facebook rules.**

Main Survey Scenarios

Each participant will randomly see half of the following 66 scenarios, and answer the three questions in Scenario Questions for each scenario.

1. Imagine you saw: **child pornography** on Facebook.
2. Imagine you saw: a photo of **someone displaying pornography to children** on Facebook.
3. Imagine you saw: a photo of **a minor in a sexual pose** on Facebook.
4. Imagine you saw: a post that **encourages people to commit suicide** on Facebook.
5. Imagine you saw: a photo of **someone committing suicide** on Facebook.
6. Imagine you saw: **an invitation to participate in terrorist activities** on Facebook.
7. Imagine you saw: **an invitation to a hate group gathering** on Facebook.
8. Imagine you saw: a post that **attempts to sell children** on Facebook.
9. Imagine you saw: a post of **a plan to kill multiple people** on Facebook.
10. Imagine you saw: a photo of **revenge porn** on Facebook.
11. Imagine you saw: a post that **threatens to show someone's revenge porn** on Facebook.
12. Imagine you saw: a photo of **someone sexually touching a drunk person** on Facebook.
13. Imagine you saw: a post that **celebrates a terrorist group** on Facebook.
14. Imagine you saw: a post that **celebrates the killing of multiple people** on Facebook.
15. Imagine you saw: a post that **celebrates a hate group** on Facebook.
16. Imagine you saw: a post that **calls for help to beat someone up** on Facebook.
17. Imagine you saw: a post that says **people from a certain country should die** on Facebook.
18. Imagine you saw: a post that **brags about hurting someone** on Facebook.
19. Imagine you saw: a post that says **people from a certain country are insects** on Facebook.

20. Imagine you saw: a post that expresses **enjoyment of other people suffering from violence** on Facebook.
21. Imagine you saw: a post that **encourages people to starve themselves** on Facebook.
22. Imagine you saw: a post that **encourages people to cut themselves** on Facebook.
23. Imagine you saw: a photo of **someone starving themselves** on Facebook.
24. Imagine you saw: a photo of **someone cutting themselves** on Facebook.
25. Imagine you saw: a post that **attempts to sell cocaine** on Facebook.
26. Imagine you saw: a post that **attempts to sell prescription drugs** on Facebook.
27. Imagine you saw: a post that **scams people into sending money** on Facebook.
28. Imagine you saw: a post of **false information about who can vote in an election** on Facebook.
29. Imagine you saw: a post that **encourages people to damage public properties** on Facebook.
30. Imagine you saw: a post that **pretends to come from a celebrity** on Facebook.
31. Imagine you saw: a post that **pretends to come from your friend** on Facebook.
32. Imagine you saw: **a secretly taken photo up someone's skirt** on Facebook.
33. Imagine you saw: a post that **harasses someone** on Facebook.
34. Imagine you saw: a post that says **people from a certain country are mentally ill** on Facebook.
35. Imagine you saw: a post that says **people from a certain country should be segregated** on Facebook.
36. Imagine you saw: a post of **racial slurs** on Facebook.
37. Imagine you saw: a post that **makes fun of other people's disabilities** on Facebook.

38. Imagine you saw: a photo of **someone beating children** on Facebook.

39. Imagine you saw: a photo of **a dismembered body on a street** on Facebook.

40. Imagine you saw: a post that **attempts to sell guns** on Facebook.

41. Imagine you saw: a post that **leaks someone's password for their bank account** on Facebook.

42. Imagine you saw: a post that **offers prostitution services** on Facebook.

43. Imagine you saw: a post that **attempts to sell marijuana** on Facebook.

44. Imagine you saw: a post of **a dismembered body in a hospital** on Facebook.

45. Imagine you saw: a photo of **someone torturing an animal** on Facebook.

46. Imagine you saw: a photo of **naked children** on Facebook.

47. Imagine you saw: a post that **asks for sex partners** on Facebook.

48. Imagine you saw: a photo of **adults having sex** on Facebook.

49. Imagine you saw: a post of **detailed description of sexual intercourse** on Facebook.

50. Imagine you saw: posts of **repeated false advertising** on Facebook.

51. Imagine you saw: a post that **claims people have to like something before they can see it** on Facebook.

52. Imagine you saw: a photo of **naked adults** on Facebook.

53. Imagine you saw: a picture of **animated pornography** on Facebook.

54. Imagine you saw: a post of **fake news** on Facebook.

55. Imagine you saw: **a link to download a pirated movie** on Facebook.

56. Imagine you saw: a post that **teaches people how to steal a car** on Facebook.

57. Imagine you saw: **an invitation to participate in a criminal group gathering** on Facebook.

58. Imagine you saw: a post that **celebrates a criminal group** on Facebook.

59. Imagine you saw: a post that **attempts to sell human organs** on Facebook.
60. Imagine you saw: a post that **attempts to sell live animals not from pet stores** on Facebook.
61. Imagine you saw: a post that **attempts to sell endangered animals** on Facebook.
62. Imagine you saw: a post that **encourages people to participate in a highly dangerous activity** on Facebook.
63. Imagine you saw: a post with **a link to a site with virus** on Facebook.
64. Imagine you saw: a post that **attempts to sell alcohol and cigarette not from a store** on Facebook.
65. Imagine you saw: a post that **misleads people about the purpose of its content** on Facebook.
66. Imagine you saw: a post that **teaches people how to hack Facebook** on Facebook.

Demographics

Lastly, we would like to know a little more about you.

What is your gender? *(Select all that apply)*

☐ Woman

☐ Man

☐ Non-binary

☐ Prefer not to disclose

☐ Prefer to self-describe: [text box]

What is your age?

- 18-24
- 25-34

- 35-44
- 45-54
- 55-64
- 65-74
- 75-84
- 85 or older

What best describes you? *(Select all that apply)*

- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Black or African American
- ☐ Native Hawaiian or Pacific Islander
- ☐ White
- ☐ Other: [text box]

What is the highest level of education you have completed?

- Less than a high school diploma
- High school diploma or GED
- Some college
- College graduate
- Some postgraduate work
- Postgraduate degree

D

APPENDIX
papers included in systematic literature review

PAPER	QUALITATIVE	QUANTITATIVE	VOLUNTEER MODERATION	COMMERCIAL MODERATION
(Rajadesingan et al., 2020)		•	•	
(Feuston et al., 2020)	•		•	•
(Fan & Zhang, 2020)	•	•		
(Juneja et al., 2020)	•	•	•	
(Hua et al., 2020)	•	•		•
(Luo et al., 2020)		•	•	•
(Phadke & Mitra, 2020)	•	•		•
(Gilbert, 2020)	•		•	
(Obar & Oeldorf-Hirsch, 2020)		•		•
(Riedl et al., 2020)		•		•
(Einwiller & Kim, 2020)	•	•		•
(Banchik, 2020)	•			•
(Schoenebeck et al., 2020)		•		•
(Skousen et al., 2020)	•		•	•
(Gray & Suzor, 2020)	•	•		•
(Datta & Adar, 2019)		•		•
(Grover & Mark, 2019)		•	•	•
(S. Jiang et al., 2019)		•		•
(Redmiles et al., 2019)	•	•		•
(Karunakaran & Ramakrishan, 2019)	•	•		•

(Kiene et al., 2019)	•		•	
(Blackwell et al., 2019)	•		•	•
(Jhaver, Bruckman, et al., 2019)		•	•	
(Jhaver, Birman, et al., 2019)	•		•	
(Jhaver, Appling, et al., 2019)	•	•	•	
(Srinivasan et al., 2019)		•	•	
(J. A. Jiang et al., 2019)	•		•	
(Chandrasekharan et al., 2019)	•	•	•	
(Wohn, 2019)	•		•	
(Fiesler & Bruckman, 2019)	•		•	
(Gibson, 2019)		•	•	
(Tyler et al., 2019)		•		•
(Potts et al., 2019)		•	•	
(Procházka, 2019)	•			•
(Squirrell, 2019)	•		•	
(Seering et al., 2019)	•		•	
(Witt et al., 2019)		•		•
(Matias, 2019)	•		•	•
(Juneström, 2019)	•	•		•
(Shen & Rose, 2019)		•		•
(Medeiros, 2019)	•			•
(Draper, 2019)	•		•	
(Suzor et al., 2019)	•			•
(Nurik, 2019)	•			•
(Duguay et al., 2018)	•			•
(Fiesler et al., 2018)	•	•	•	•
(Blackwell et al., 2018)	•	•	•	
(Jhaver et al., 2018)	•	•	•	
(Matias & Mou, 2018)	•		•	
(Chancellor et al., 2018)		•	•	
(Pavalanathan et al., 2018)		•	•	
(Chandrasekharan et al., 2018)	•	•	•	
(Gerrard, 2018)	•			•
(West, 2018)	•	•		•
(Keegan & Fiesler, 2017)		•	•	
(Chancellor et al., 2017)		•		•

(Pellicone & Ahn, 2017)	•		•	
(Blackwell et al., 2017)	•			•
(Chandrasekharan et al., 2017)		•		•
(Seering et al., 2017)		•	•	
(Zeng et al., 2017)		•		•
(Cheng et al., 2017)		•		•
(Newell et al., 2016)	•	•		•
(Chancellor, Pater, et al., 2016)		•		•
(Park et al., 2016)	•			•
(Centivany & Glushko, 2016)	•		•	
(Matias, 2016a)	•	•		•
(Gallagher & Savage, 2016)		•		•
(Getto & Labriola, 2016)	•		•	
(Kiene et al., 2016)	•		•	
(Ehrett, 2016)	•	•	•	•
(Vashistha et al., 2015)	•	•		•
(Wang et al., 2014)		•		•
(Petrič & Petrovčič, 2014)		•	•	•
(Lampe et al., 2014)	•	•	•	
(Kayhan & Bhattacherjee, 2013)		•		•
(Heinze et al., 2013)	•	•		•
(Sarkar et al., 2012)		•		•
(Holmes & Cox, 2011)	•	•	•	
(Liao et al., 2010)	•	•	•	•
(Gurzick et al., 2009)	•			•
(Lampe & Johnston, 2005)		•	•	•
(Lampe & Resnick, 2004)	•	•	•	•

Table D-1. Papers included in systematic literature review.