Identifying and Addressing Design and Policy Challenges in Online Content Moderation

Jialun "Aaron" Jiang
University of Colorado Boulder
Boulder, CO 80309, USA
aaron.jiang@colorado.edu

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.
Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6819-3/20/04.
http://dx.doi.org/10.1145/3334480.3375030

Abstract
My doctoral research identifies and addresses the design and policy challenges of content moderation on the levels of smaller communities and larger platforms, using a combination of interviews and surveys. My prior work studies voice-based communities on Discord, and shows that voice as a new technology brings unique challenges to community moderators and highlights the fundamental evidentiary need in content moderation. My dissertation expands on this work by (1) quantifying severity of violations as a way to address social media platforms' challenge of prioritization, and (2) empirically evaluating existing moderation approaches and design recommendations by soliciting moderator and user perceptions of them. Combining these perspectives, my dissertation will provide guidelines for improving content moderation in online communities.

Author Keywords
Content moderation; regulation; online communities; social media

CCS Concepts
•**Human-centered computing** → **Empirical studies in collaborative and social computing;** *Social media;* Social networking sites;

Context and Motivation

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 1998 book *My Tiny Life* [3], he describes "a rape in cyberspace" in which Mr. Bungle, a user of the text-based community LambdaMOO, forced 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, nor did LambdaMOO 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 people getting sexually harassed and attacked in virtual reality. It is clear that online communities today are still facing similar, but more complicated and challenging problems.

As both platforms and users clamor for solutions to these problems, one often put forth is content moderation, which James Grimmelmann defines as the "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 subreddits with a few thousand members. But 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 possibly violating content when they are flagged, but many volunteer moderators also devote their time to combating these problems in their own, smaller communities on a daily basis. We have seen many success stories of moderating online communities, ranging from using Reddit AutoModerator to automatically remove problematic content [12, 10], to setting positive examples to encourage similar behaviors in Twitch chat [15]. On the platform level, prior research also shows that banning hate communities reduces the amount of overall hate speech Moreover, when and the communities that the hate community members migrated to did not inherit their behavior [2].

However, we have not yet heard the good news that all the problems in online communities have been solved—content moderation work is not perfect, and it is still faced with many challenges. Not only do they have to do the often thankless job of continuously looking at traumatic content and making personally uncomfortable decisions [18, 4], moderators also often had to resort to imperfect solutions to problems in their communities. For example, while having clear, prominently displayed rules is helpful for community members to learn the norms, it may convey the message that these rules are often broken [13], and make community members feel stifled and constrained [12]. Furthermore, the limited availability of volunteer community moderators means that moderation is often delayed [14], leaving problematic content in place, and community members' effort to circumvent moderation [6, 1] makes timely moderation even harder.

While there has been an emergence of machine-learning based automated moderation tools, it is difficult to gather enough training data for rule violations when rule breakers try to hide themselves—the classic unresolved machine

learning problem of imbalanced classification, and automated moderation tools are not adaptive to new kinds of expression and communication [7]. Even with the help of automated moderation tools, moderators still need to make nuanced, case-by-case punishment decisions that automated moderation is not capable of [16], but how to scale this human labor is still an open and challenging question with today's incredibly complex online communities with millions of users of diverse backgrounds, purposes, and values [7]. These regulation challenges not only speak to the design of technology, but are deeply connected to platform policies as well, and my dissertation aims to identify and address the design and policy challenges in online content moderation.

Research Questions

On a high level, my dissertation work is guided by the following research questions:

1. What are some of the design and policy challenges in content moderation?
2. What actions can we take to address these challenges to improve moderation?
3. How useful are these recommended actions for real moderators and users?

Prior Work

My prior work in content moderation begins with a large-scale analysis of the rules of 100,000 Reddit subcommunities ("subreddits") [5]. Through a combination of qualitative content analysis and computational analysis, I found that only 52% of subreddits have any rules, and the majority of rules are restrictive (i.e., "Don't do this.") rather than prescriptive (i.e., "Do this."), and rarely do they mention higher-level platform guidelines or policies. The restrictive nature of rules suggest the possibility that rules are mostly reactive to emergent misbehaviors and are not comprehensive. These lack of rules in many communities also suggest that there might be challenges in creating effective governance structures to start with.

My completed work for my dissertation includes a study that investigates the challenges that a new technology in online communities—real-time voice—presents to volunteer moderators [11]. Through in-depth semi-structured interviews with 25 moderators on Discord, a social platform with real-time voice chat being a primary mode of communication, I found that voice enabled new ways to break rules that the moderators did not expect, and the complexity of misbehaviors in voice made it difficult for moderators to create specific rules to prevent such behaviors. Furthermore, the ephemerality of real-time voice also highlights the fundamental evidentiary need of content moderation that did not show up in existing moderation literature that focuses on text-based online spaces. While Discord moderators sometimes resorted to recording to acquire evidence, such action is against the law in some regions, and the vague platform policy offloaded the responsibility to discern different laws in the world to the moderators.

This study contributes an understanding the design and policy challenges that platform technology presents, and points to ways that technology designers and policy makers can support the moderators who face these challenges everyday. Following this work, I shift my attention to addressing one of most pressing policy challenges in platform moderation—prioritization of moderation—and then empirically examine and evaluate the practical values of the

potential solutions and design recommendations that researchers have made.

Future Dissertation Work

My proposed dissertation work will include two more studies. The first study will address social media platforms' challenge prioritization of moderation work by revealing users' perceptions of the severity of a wide range of rule violations and misbehaviors; the second study will empirically examine the usefulness and implications of the design recommendations related to content moderation in prior literature.

Study 1: Quantifying Severity of Rule Violations on Social Media

There are many kinds of rule violations on social media, but not all of them were created equal. Intuitively, we know that promoting terrorism is worse than spamming, but *how much* worse is it? And how should platforms allocate moderation capacity across all the possible violations? Answers to these questions not only provide guidelines for social media platforms that constantly need to prioritize resolving some violations over others, but also improve our understanding of misbehaviors online.

In this study, I will use a survey that asks users to rate the severity of all violations listed in Facebook's Community Standards. 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 responses options are equidistant and the tendency to cause anchor effects [9]. Using a bounded scale that typically has five or seven options also limits our ability to model an accurate relationship between more than fifty violations. Therefore, I will use a novel method that asks participants to directly respond with a number as free text that represents their perceived severity of a violation.

As severity of rule violations can be an unfamiliar concept to regular users, in the survey I will operationalize severity in two ways: punishment and urgency. Punishment roots from the concept of *proportionality* [17] in criminal justice literature: the more severe a violation is, the graver the associated punishment should be. The construct of urgency is inspired by emergency response systems: the more severe a violation is, the sooner it should be taken care of.

Because the severity of violations can be culturally-dependent, I will strategically recruit survey participants from countries covering North America, South America, Asia-Pacific (APAC), and Europe, the Middle East and Africa (EMEA). Having a broad range of participants from different cultures will help me analyze the similarities and differences in perceptions between cultures. I will use statistical methods to analyze the survey responses and the results will include localized priority of violations, statistical models of the relationship between violations, and similarities and differences between these models in different cultures.

Study 2: Moderator and User Perceptions of Moderation Design Recommendations

Having implications and recommendations for design is a common practice in HCI scholarship. While there has been an abundance of design recommendations related to content moderation made in the HCI literature, rarely do researchers know how these design recommendations are viewed by their beneficiaries—the communities, and how they may function differently in different communities.

To address this gap, in this study, I will first conduct a systematic literature review of prior work related to online content moderation. Then I will directly take the design rec-

ommendations identified in the literature review, and ask moderators and users about their perceptions and opinions of these recommendations, including their perceived usefulness, potential impact, and potential challenges of implementation. I expect the findings from this survey will spark conversations in the moderation research community, and provide guidelines for researchers to better serve communities.

Expected Contributions

I expect my dissertation to offer the following contributions:

- Through my prior work with moderators of voice-based communities on Discord, this dissertation identifies new types of rules and new ways to break these rules that do not exist in text-based communities, as well as moderators' struggle to handle them. The findings from this study also challenge and complicate Grimmelmann's framework of moderation techniques [8], and highlight the fundamental evidentiary need in content moderation that will apply not only in voice-based communities. By doing so, this dissertation reveals design challenges in content moderation in a new technology, and offer design and policy recommendations to address these challenges.
- Through a proposed study that investigates the severity of violations (Study 1), my dissertation will address a policy challenge of content moderation by providing empirical evidence of how different kinds of violations can be prioritized. Having established quantified severity, my dissertation will also point to ways to mitigate certain contents' negative impact to moderators, as severe violations are often associated with intense content (e.g., graphic violence). By using questions that take free-form and unbounded responses, my dissertation also contributes a novel method to measure perceptions and opinions as an alternative to the conventional Likert scale.
- Through a proposed study that solicits moderator and user perceptions of design recommendations for content moderation (Study 2), this dissertation will contribute a deep understanding and critical inquiry of the practical values of the common practice of design recommendations in HCI and social computing research in the context of content moderation.

Together, this dissertation aims to provide a comprehensive overview of the challenges in moderation and regulation in online spaces. As online malicious behavior becomes increasingly rampant, it is critical for us to examine where content moderation currently falls short, and identify actions that we can take to make our communities safer.

REFERENCES

[1] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1201–1213. DOI: http://dx.doi.org/10.1145/2818048.2819963

[2] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.: CSCW* 1, 2 (2017). DOI: http://dx.doi.org/10.1145/3134666

[3] Julian Dibbell. 1998. *My Tiny Life: Crime and Passion in a Virtual World*. Julian Dibbell.

[4] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of 2019 ACM Conference on Human Factors in Computing Systems (CHI'19)*. ACM, Glasgow, UK.

[5] Casey Fiesler, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Twelfth International AAAI Conference on Web and Social Media*. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17898

[6] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (Dec. 2018), 4492–4511. DOI:http://dx.doi.org/10.1177/1461444818776611

[7] Tarleton Gillespie. 2018. *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, New Haven.

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

[9] J. P. Guilford. 1954. *Psychometric methods, 2nd ed*. McGraw-Hill, New York, NY, US.

[10] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5 (July 2019), Article 31. DOI: http://dx.doi.org/10.1145/3338243

[11] Jialun "Aaron" Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), Article 55. DOI: http://dx.doi.org/10.1145/3359157

[12] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1152–1156. DOI:http://dx.doi.org/10.1145/2858036.2858356

[13] Robert E. Kraut and Paul Resnick. 2011. *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press, Boston, MA.

[14] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 543–550. DOI: http://dx.doi.org/10.1145/985692.985761

[15] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 111–125. DOI: http://dx.doi.org/10.1145/2998181.2998277

[16] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (Jan. 2019), 1461444818821316. DOI:http://dx.doi.org/10.1177/1461444818821316

[17] Andrew von Hirsch. 1992. Proportionality in the Philosophy of Punishment. *Crime and Justice* 16 (1992), 55–98. https://www.jstor.org/stable/1147561

[18] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How they Get Involved, the Roles they Play, and the Emotional Labor they Experience.. In *Proceedings of 2019 ACM Conference on Human Factors in Computing Systems (CHI'19)*. ACM, New York, NY, USA.