ABSTRACT

UNDERSTANDING THE VOLUNTARY MODERATION PRACTICES IN LIVE STREAMING COMMUNITIES

by
Jie Cai

Harmful content, such as hate speech, online abuses, harassment, and cyberbullying, proliferates across various online communities. Live streaming as a novel online community provides ways for thousands of users (viewers) to entertain and engage with a broadcaster (streamer) in real-time in the chatroom. While the streamer has the camera on and the screen shared, tens of thousands of viewers are watching and messaging in real-time, resulting in concerns about harassment and cyberbullying. To regulate harmful content—toxic messages in the chatroom, streamers rely on a combination of automated tools and volunteer human moderators (mods) to block users or remove content, which is termed *content moderation*. Live streaming as a mixed media contains some unique attributes such as synchronicity and authenticity, making real-time content moderation challenging.

Given the high interactivity and ephemerality of live text-based communication in the chatroom, mods have to make decisions with time constraints and little instruction, suffering cognitive overload and emotional toll. While much previous work has focused on moderation in asynchronous online communities and social media platforms, very little is known about human moderation in synchronous online communities with live interaction among users in a timely manner. It is necessary to understand mods' moderation practices in live streaming communities, considering their roles to support community growth. This dissertation centers on volunteer mods in live streaming communities to explore their moderation practices and relationships with streamers and viewers. Through quantitative and qualitative methods, this dissertation mainly focuses on three aspects: the strategies and tools used by moderators, the mental model and decision-making process applied toward violators, and the conflict management present in the moderation team. This dissertation uses various socio-technical theories to explain mods' individual and collaborative practices and suggests several design interventions to facilitate the moderation process in live streaming communities.

UNDERSTANDING THE VOLUNTARY MODERATION PRACTICES IN LIVE STREAMING COMMUNITIES

by
Jie Cai

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Information Systems

Department of Informatics

May 2022

APPROVAL PAGE

UNDERSTANDING THE VOLUNTARY MODERATION PRACTICES IN LIVE STREAMING COMMUNITIES

Jie Cai

Dr. Donghee Y. Wohn, Dissertation Advisor — Date
Associate Professor, New Jersey Institute of Technology

Dr. Aritra Dasgupta, Committee Member — Date
Assistant Professor, New Jersey Institute of Technology

Dr. David Wang, Committee Member — Date
Assistant Professor, New Jersey Institute of Technology

Dr. Cody Buntain, Committee Member — Date
Assistant Professor, University of Maryland, College Park, MD

Dr. Bryan Dosono, Committee Member — Date
Senior Design Researcher, eBay Inc.

BIOGRAPHICAL SKETCH

Author:	Jie Cai
Degree:	Doctor of Philosophy
Date:	May 2022

Undergraduate and Graduate Education:

- Doctor of Philosophy in Information Systems,
  New Jersey Institute of Technology, Newark, NJ, 2022
- Master of Science in Marketing Research,
  Hofstra University, Long Island, NY, 2016
- Bachelor of Administration in Marketing/Organizational Leadership (dual degree),
  Shenyang Normal University/Fort Hays State University, Shenyang, China, 2014

Major:	Information Systems

Presentations and Publications:

Jie Cai and Donghee Yvette Wohn, "Coordination and Collaboration: How do Volunteer Moderators Work as a Team in Live Streaming Communities?" In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1-14, 2022.

Christine Cook, **Jie Cai**, and Donghee Yvette Wohn, "Awe Versus Aww: The Effectiveness of Two Kinds of Positive Emotional Stimulation on Stress Reduction for Online Content Moderators," In *Proceedings of the ACM on Human-Computer Interaction*,(CSCW):1-19, 2022.

Sahaj Vaidya, **Jie Cai**, Soumyadeep Basu, Azadeh Naderi, Donghee Yvette Wohn, and Aritra Dasgupta, "Conceptualizing Visual Analytic Interventions for Content Moderation," In *Proceedings of the IEEE Visualization Conference*, pages 1-5, 2021.

Jie Cai and Donghee Yvette Wohn, "After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities," In *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.

Jie Cai, Ruiqi Shen, Starr Roxanne Hiltz, "Understanding Choice of Social Music Systems in China: A Study of NetEase Cloud Music," In *Adjunct Publication of the 23rd International Conference on Mobile Human-Computer Interaction*, pages 1-6, 2021.

Jie Cai, Sarah J Ryu, Donghee Yvette Wohn, and Hyejin Hannah Kum-Biocca, "Teleworker's Perception of Technology Use for Collaborative and Social During the COVID-19 Pandemic," In *Proceedings of the International BCS Human Computer Interaction Conference*, pages 1-14, 2021.

Jie Cai, Cameron Guanlao, and Donghee Yvette Wohn, "Understanding Rules in Live Streaming Micro Communities on Twitch," In *Proceedings of the ACM International Conference on Interactive Media Experiences*, pages 290-295, 2021.

Jie Cai and Donghee Yvette Wohn, and Mashael Almoqbel, "Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities," In *Proceedings of the ACM International Conference on Interactive Media Experiences*, pages 61-72 2021.

Jirassaya Uttarapong, **Jie Cai**, and Donghee Yvette Wohn, "Harassment Experiences of Women and LGBTQ Live Streamers and How They Handled Negativity," In *Proceedings of the ACM International Conference on Interactive Media Experiences*, pages 7-19, 2021.

Jie Cai and Donghee Yvette Wohn, "Categorizing Live Streaming Moderation Tools: An Analysis of Twitch," *International Journal of Interactive Communication Systems and Technologies*, 9(2):36-50, 2019.

Jie Cai and Donghee Yvette Wohn, "Live Streaming Commerce: Uses and gratifications Approach to Understanding Consumers' Motivations," In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2548-2557, 2019.

Jie Cai and Donghee Yvette Wohn, "What are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives," In *Conference Companion Publication of the ACM on Computer Supported Cooperative Work and Social Computing*, pages 166-170, 2019.

Jie Cai, Donghee Yvette Wohn, and Guo Freeman, "Who Purchases and Why? Explaining Motivations for In-game Purchasing in the Online Survival Game Fortnite," In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 391-396, 2019.

Jie Cai, Ankit Mittal, Dhanush Sureshbabu, and Donghee Yvette Wohn, "Utilitarian and Hedonic Motivations for Live Streaming Shopping," In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, pages 81-88, 2018.

献给我的祖国，父母，家人，还有即将出生的蔡含章。

蔡杰

To my motherland, parents, family, and upcoming son Hanzhang Arthur Cai.

Jie Cai

ACKNOWLEDGMENT

I would like to thank NJIT for giving me the opportunity to pursue a Ph.D. I still remember the moment that I received the offer letter in July 2017 and joined the program two months later.

I also want to thank my advisor, Dr. Donghee Yvette Wohn, who keeps pushing and advising me to learn everything that I can never learn from the class. I feel I am so lucky to have you be my advisor from my heart. You are an excellent collaborator, mentor, and friend, to teach me, advise me, share your academic experience, respect my decision, etc. I can't achieve success without your support.

Thank all to my committee members (Dr. Aritra Dasgupta, Dr. Cody Buntain, Dr. David Wang, Dr. Bryan Dosono) for serving your time to support me achieving this milestone in my life. I also want to thank Dr. Starr Roxanne Hiltz for being my dissertation proposal committee member.

My research is fortunately supported by two National Science Foundation grants (Award Nos. 1841354, 1928627).

I want to thank the undergraduate research assistants from the Social Interaction Lab (Jessy Martinez, Aaron Samuel, Andrew Suarez, Abdelmalek Benaissa, etc.) for helping with participant recruitment, video analysis, transcription, and coding. I really appreciate undergraduates' support of my research and would like to keep working with you guys after my graduation. You are so talented and innovative.

I want to thank my wife, Jiali Qi, who always accompanies and supports me when I face much pressure at the beginning of the Ph.D. journey.

TABLE OF CONTENTS

Chapter Page

1 INTRODUCTION . . . 1

1.1 Motivation and Research Problems . . . 3

1.2 Approach . . . 4

1.3 Dissertation Overview . . . 5

2 BACKGROUND . . . 8

2.1 Community Moderation, Human Moderator, Moderation System . . . 8

2.2 Moderation in Live Streaming Community and Twitch . . . 10

2.3 Conclusion . . . 12

3 MODERATION VISIBILITY: MAPPING THE STRATEGIES OF VOLUNTEER MODERATORS IN LIVE STREAMING MICRO COMMUNITIES . . . 13

3.1 Introduction . . . 13

3.2 Methods . . . 14

3.2.1 Participant Recruitment . . . 14

3.2.2 Interview . . . 15

3.2.3 Participant Demographics . . . 16

3.3 Results . . . 17

3.3.1 Proactive Strategies . . . 18

3.3.2 Reactive Strategies . . . 22

3.4 Discussion . . . 33

3.4.1 Interactivity Facilitates the Visible and Performative Activities of Moderation . . . 33

3.4.2 Synchronicity Enhances the Graduated Moderation and Amplifies the Violator's Voice . . . 35

3.4.3 Design Implications . . . 37

3.4.4 Limitations . . . 38

3.5 Conclusion . . . 39

TABLE OF CONTENTS
(Continued)

Chapter **Page**

4 CATEGORIZING LIVE STREAMING MODERATION TOOLS: AN ANALYSIS OF TWITCH . . . 40

4.1 Introduction . . . 40

4.2 Results . . . 40

4.2.1 Moderation Tools . . . 40

4.2.2 The Desired New Features . . . 45

4.3 Discussion . . . 49

4.3.1 Design Opportunities . . . 50

4.3.2 Suggested New Functions . . . 51

4.3.3 Limitations and Future Work . . . 52

4.4 Conclusion . . . 53

5 WHAT ARE EFFECTIVE STRATEGIES OF HANDLING HARASSMENT ON TWITCH? USERS' PERSPECTIVES . . . 54

5.1 Introduction . . . 54

5.2 Methods . . . 54

5.3 Results . . . 55

5.4 Discussion . . . 59

5.5 Conclusion . . . 60

6 AFTER VIOLATION BUT BEFORE SANCTION: UNDERSTANDING VOLUNTEER MODERATORS' PROFILING PROCESSES TOWARD VIOLATORS IN LIVE STREAMING COMMUNITIES . . . 61

6.1 Introduction . . . 61

6.2 Related Work . . . 62

6.2.1 Criminal Profiling . . . 62

6.2.2 Applying Criminal Profiling to Community Moderation and in Live Streaming Communities . . . 64

6.3 Methods . . . 66

6.3.1 Participant Recruitment and Demographics . . . 67

TABLE OF CONTENTS
(Continued)

Chapter **Page**

6.3.2 Video Analysis and Interview Process 67

6.3.3 Interview Analysis 70

6.4 Results 70

6.4.1 Evidence Types 71

6.4.2 Evidence Collection 80

6.4.3 Types of Violators 84

6.5 Discussion 89

6.5.1 Platform Design and Affordance Make Profiling Go Beyond the User's Profile 90

6.5.2 Profiling as Part of the Moderation for Community Growth . . 91

6.5.3 Implications and Recommendations 93

6.5.4 Limitation and Future Work 96

6.6 Conclusion 96

7 CONFLICT WITH MANAGEMENT IN THE MODERATION TEAM IN LIVE STREAMING COMMUNITIES 98

7.1 Introduction 98

7.2 Related Work 99

7.2.1 Organizational Conflict and Conflict in Online Communities . 100

7.2.2 Online Community Commitment 101

7.2.3 Conflict Management in Online Communities 103

7.3 Methods 104

7.3.1 Participant Recruitment 105

7.3.2 Participant Demographic 106

7.3.3 Survey Measure 106

7.3.4 Open-ended Question Analysis 107

7.4 Results 108

7.4.1 Descriptive Results 108

TABLE OF CONTENTS
(Continued)

Chapter **Page**

7.4.2 Hypotheses Test 109

7.4.3 RQ1a: Relationship Between Conflict Types and Conflict Management Styles 110

7.4.4 RQ1b: Incidents of Conflict and Conflict Management Styles . 110

7.4.5 RQ2: Commitment to the Streamer and Conflict Management Styles 119

7.5 Discussion 120

7.5.1 Conflicts in the Team and Commitments to the Streamer Are Independent 121

7.5.2 Active and Cooperative Style Versus Passive and Assertive Style 122

7.5.3 High Concerns for the Streamer or Low Concerns for Mods Themselves 124

7.5.4 Design Implications 125

7.5.5 Limitations and Future Work 126

7.6 Conclusion 127

8 CONCLUSION 129

REFERENCES 131

LIST OF TABLES

Table **Page**

3.1 Moderator Demographics and Moderation Activity 17

4.1 Moderation Tool Categories and Functions 41

5.1 Difference Test 55

5.2 Exploratory Factor Analysis of Effective Strategies 57

5.3 Other Strategies to Combat Toxicity 58

5.4 Quotes of Other Strategies 59

6.1 Demographic and Experience of Participants 68

7.1 H1,2,3: Correlations among Conflicts and Commitments to the Streamer 109

7.2 RQ1a: Correlations among Conflicts and Management Styles 110

7.3 RQ2: Regression Model Examining the Effect of Commitments to the Streamer on Conflict Management Styles 120

LIST OF FIGURES

Figure **Page**

1.1 A screenshot of the Twitch interface: the content producer is streaming content on the left side of the screen; viewers are commenting in the highlighted chatroom on the right side. 2

2.1 Twitch chatroom interface from viewers' view (Left, colorful usernames with different badges to indicate status) and mods' view (Right, shortcuts of "Ban" "Timeout" "Delete" are visible next to the usernames). 11

3.1 Moderation strategies for before and after a bad act happens. Lines indicate relationships, arrows indicate sequence. 18

5.1 Who should be responsible for deciding how to enforce what is appropriate? 56

6.1 Screenshots of the interface from a moderator's view integrated with different moderation tools. (a) Box 1: the LeaderBoard; Box 2: badges; the icons alongside each blurred usernames are three commands: ban, timeout, delete. (b) A moderator checks a user's message history: this user sent only 5 messages in the chat with 0 timeout, bans, and mod comments. 74

6.2 A message deleted, and a user timed out by P18. The icons alongside the blurred usernames are commands without any badges, such as ban, two different timeouts, delete, etc. 75

6.3 A screenshot of Whisper conversation between a viewer and a mod on Twitch. After a violator apologized, the mod unbanned them with an explanation and maintained the "violator" in the community. 78

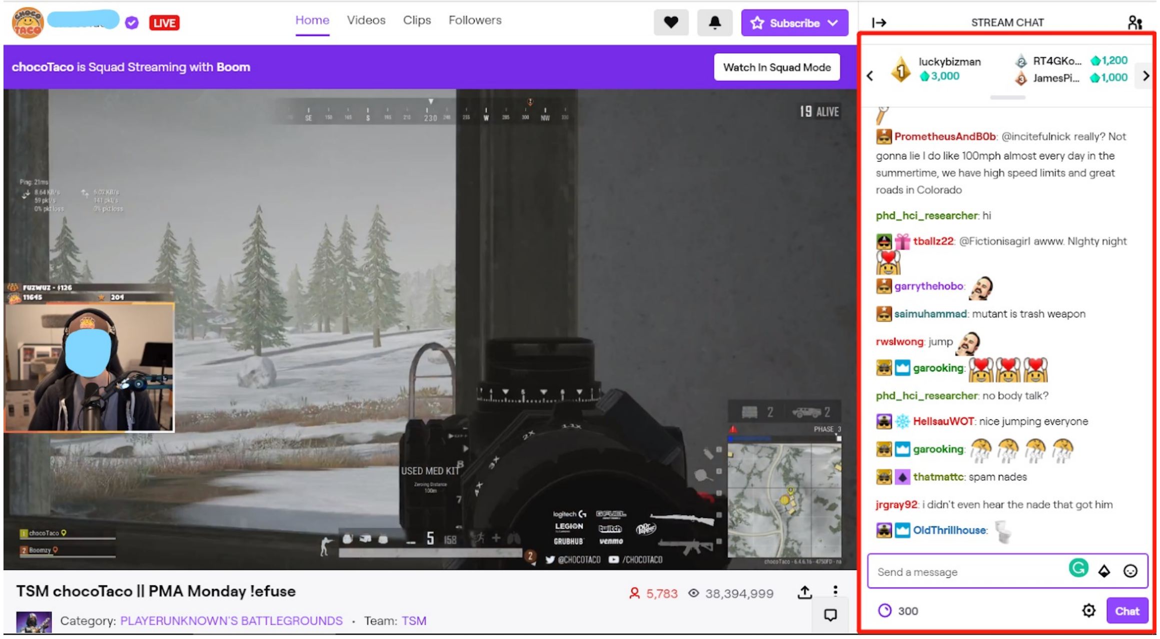
CHAPTER 1

INTRODUCTION

Online abuse, such as hateful speech, sexual harassment, personal attack, and doxing, is a severe and pervasive social problem. According to a research survey by the Anti-Defamation League, 44% of Americans report that they experience online harassment. In some cases, these experiences are coupled with other impacts, such as anxiety and thoughts of depression and suicide [101].

In order to reduce online abuse and maintain the growth and health of online communities, commercial platforms apply many techniques to filter abusive language, such as improving algorithms and applying tools (e.g., [85, 87, 24, 10]). However, violators always seek ways to circumvent the algorithms and cheat the tools with variants [55, 23]. To supplement algorithmic moderation, platforms also rely on human moderators (mods), either active volunteer users [133] or well-trained content experts [124], to manually remove user-reported content or review incidents in context-sensitive situations [129]. With the adoption of new technology and the evolution of communities, moderation practice faces new challenges.

Live streaming is a rapidly growing industry, estimated to reach 70.5 billion USD by the year 2021 [109]. Acquired by Amazon in 2014, Twitch started live streaming services very early by focusing on the gaming genre and has become one of the global leading live streaming platforms. Now it is broadening into many other categories including IRL (in real life), creative, food & drink, and travel & outdoors. Live streaming enables the streamer to share the rich ephemeral experience with informal social interaction with viewers in the chat [68]. The broadcasting element enables the streamer to transmit content to many viewers, and real-time Internet Relay Chat enables viewers to comment and interact with the streamer in the chatroom [159].



PrometheusAndB0b: @incitefulnick really? Not gonna lie I do like 100mph almost every day in the summertime, we have high speed limits and great roads in Colorado

phd_hci_researcher: hi

tballz22: @Fictionisagirl awww. NIghty night

garrythehobo:

saimuhammad: mutant is trash weapon

rwslwong: jump

garooking:

phd_hci_researcher: no body talk?

HellsauWOT: nice jumping everyone

garooking:

thatmattc: spam nades

jrgray92: i didn't even hear the nade that got him

OldThrillhouse:

Figure 1.1 A screenshot of the Twitch interface: the content producer is streaming content on the left side of the screen; viewers are commenting in the highlighted chatroom on the right side.

Figure 1.1 shows the interface of Twitch, a typical live streaming platform. Such a new mode of interaction and adoption of live broadcasting technology introduces new ways of communication and content creation, norm violation, and consequently, content moderation. The synchronicity and ephemerality of live streaming render different challenges for the hidden labor of human moderators compared with those of other online communities. Like other live streaming services, Twitch requires a high need for real-time moderation. In this dissertation, content moderation focuses on the text messages created by viewers in the chatroom, not the broadcasting content created by the streamer.

Unlike other social media such as Facebook or YouTube, where moderators are appointed by the company to review content that is reported by users [124], live streaming communities heavily rely on volunteer moderators ("mods"). Nevertheless, little is known about how they moderate their channels and how they interact with viewers and communicate with the streamer, a gap that my dissertation aims to fill.

1.1 Motivation and Research Problems

The popularity of live streaming and the success of Twitch have made it a growing subject of academic attention. Most current research on live streaming, however, focuses on streamers and viewers, such as streamer or viewer motives [16, 20, 53, 126] and streamer-viewer interactions [106, 159], with less but growing attention on the prominent but hidden role of human moderators [132, 157].

Ample work has explored the algorithms and moderation tools to automatically prevent people from violating the rules and to detect and remove the harmful content at scale [59, 23], but moderation only by advanced technologies still has some limitations, such as failing to understand the context and encouraging deviant behaviors [23]. Thus, human moderators are needed for moderation scenarios where algorithms have limited efficacy. Nonetheless, content creation is growing at an exponential speed and exceeding the capability of current human moderators. Different from commercial companies' strategies to handle harmful content by mainly hiring people [124], many user-governed online communities such as Reddit and Twitch, which contain diverse micro-communities, encourage volunteer moderators to moderate their micro-communities.

The high interactivity of text-based communities prompts a large volume of messages dynamically flowing in the chat and disappearing quickly. Mods need immediate attention to these messages; high concentration with the time constraint of this kind of information causes information overload and emotional toll [157]. Thus, understanding moderation practices is crucial for us to identify the challenges mods face and to provide possible design interventions to support them.

The overall objective of my research is to understand volunteer moderators' relationships with viewers and the streamer, to identify the challenges they face during the moderation process, and to provide possible social and technical interventions to increase moderation efficiency and maintain the community with less punitive and more accurate moderation. In order to achieve my objectives, this dissertation aimed to understand mods' practices regarding strategies and tools to deal with viewers in their decision-making process toward violators, and conflict management in the moderation team with three high-level questions.

- What are the strategies and tools that volunteer mods use in live streaming communities?
- How do volunteer mods profile violators in live streaming communities?
- What do volunteer mods do to manage conflicts in the moderation team in live streaming communities?

1.2 Approach

My dissertation is conducted in three phases. In the initial phase, I have completed 21 semi-structured interviews with Twitch moderators that were recruited through several different approaches such as personal contact and posting on Twitter through an official lab account. I offered a $20 Amazon gift card for their voluntary participation. Thematic analysis [13] was used during the analysis to code answers into concepts and group the relevant concepts into themes.

In the second phase, I applied mixed methods (observation + interview) with another 19 Twitch mods. I developed a new interview protocol based on some of the findings from the first phase and plan to gain more insights into some unique findings. In addition to a semi-structured interview, we decided to add observation so that we can take the specific context of moderation into consideration. We asked the moderators to record their moderation screens for an hour and send them to us for a review first. Next, we scheduled the semi-structured interview. Each participant spent one hour on video recording and another 40-60 minutes for the interview. We offered a $100 digital Amazon gift card to each participant. I also validated moderation strategies in a quantitative way; I designed a paper survey based on some insights of interview study in the first phase and collected data from Twitch Convention 2018.

In the third phase, I aimed to understand the relationships of mods' perceptions among conflicts, conflict management styles, and commitments to the streamer, based on the findings of the first-phase study. I designed a survey to collect 240 qualified self-reported data from mods. I used a recruitment platform called Prolific[1] to collect the data. The platform used its user pool and automatically matched and distributed the survey to potential targets based on users' self-reported information on its platform. This survey took about 10-15 minutes to complete with $2 for compensation.

1.3 Dissertation Overview

Chapter 2 positions this in the context of related research into moderation in online communities. Chapters 3 and 4 and describe the results in the first phase and answer certain fundamental questions about moderation in live streaming communities such as moderation tools, moderation strategies, and moderation guidelines. Chapter 3 is under revision in *"ACM IMX 2021"* and Chapter 4 has been published in the *"International Journal of Interactive Communication Systems and Technologies."* Chapter 5 and 6 describe the results in the second phase. Specifically, Chapter 5 discusses a survey study based on the findings in the first phase about the responsibility of roles and moderation strategies in live streaming communities, to some extent, validating the qualitative results in quantitative ways. The results have been published in *"CSCW 2019."* Chapter 6 investigates the profiling process based on the profiling strategy in the first phase and aims to systematically understand how this pre-moderation strategy works. The results have been published in *"CSCW 2021."* Chapter 7 used the conflict management framework and commitment model to

[1] `https://prolific.co/`. Retrieved on March 14, 2022

explore relationship management in the moderation team (in submission). Chapter 8 was the conclusion of the dissertation.

- In Chapter 3, I mapped out 13 moderation strategies and presented them in relation to the bad act, enabling us to categorize from proactive and reactive perspectives and identify communicative and technical interventions. This study found that the act of moderation involves highly visible and performative activities in the chat and invisible activities involving coordination and sanction. The juxtaposition of real-time individual decision-making with collaborative discussions and the dual nature of visible and invisible activities of moderators provide a unique lens into a role that relies heavily on both the social and technical.

- In Chapter 4, I categorized the current features of real-time moderation tools on Twitch into four functions (chat control, content control, viewer control, settings control) and explored some new features of tools that they wish to own (e.g., grouping chat by languages, pop-out window to hold messages, chat slow down, a set of buttons with pre-written/pre-message content, viewer activity tracking, all in one).

- In Chapter 5, I surveyed 375 Twitch users in person at Twitch Convention 2019, asking them about who should be responsible for deciding what should be allowed and what strategies they perceived to be effective in handling harassment. This study found that users think that streamers should be most responsible for enforcing rules and that either blocking bad actors, ignoring them, or trying to educate them are the most effective strategies.

- In Chapter 6, I interviewed 19 Twitch moderators with 10 observations. I applied the psychological profiling model to understand how moderators profile violators before the moderation. This study found that mods engage in a complex process of collaborative evidence collection and profile violators into different categories to decide the type and extent of punishment. Mods consider violators' characteristics as well as behavioral history and violation context before taking moderation action. The main purpose of the profiling was to avoid excessive punishment and aim to integrate violators more into the community.

- In Chapter 7, I conducted an online survey (N= 240) with live streaming mods to explore their commitment to the streamer to grow the micro community and the different ways in which they handle conflicts with other mods and the streamer. I found that 1) conflicts in the team and commitments to the streamer are generally independent, though normative conflict is positively but weakly related to normative commitment to the streamer; 2) active and cooperative styles are more effective than passive and assertive styles for mods to manage conflicts, but they might be forced to do so; 3) mods with strong commitments to the streamer would like to apply styles showing either high concerns for the streamer or low concerns for themselves to manage conflicts.

CHAPTER 2

BACKGROUND

2.1 Community Moderation, Human Moderator, Moderation System

Moderation refers to *"the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse"* [63] and is the gateway for online communities to thrive as harassment, trolls, and hate speech are increasing in these spaces [12], ranging broadly from learning communities (e.g., [160, 134]) to crowd-sourcing communities (e.g., [33, 28]) to social media platforms (e.g., [124, 133, 37, 87]).

Social media platforms employ a large group of commercial content moderators (mods) or freelancers who work on contract with them [124] to supplement algorithmic moderation at scale [59]. Mods are gatekeepers [124] of commercial platforms to maintain the community health and growth [138] with the power to remove harmful content and sanction users posting the content, namely violators. However, abusing moderation power or overly sanctioning users could deter community engagement and alienate community members [138]. Mods have to trade off the punishment efficacy with community growth [28, 138]. In addition, platforms rely on users' reports who flag the potentially offensive content and then ask the moderator to review and remove the content manually [55, 35]. Users can also apply tools such as 'Blocklist' on Twitter to block harassers [87].

Different from most social media platforms that handle moderation within, user-governed communities such as Wikipedia and Reddit rely on volunteer moderators who are given limited administrative power to remove unacceptable content and ban violators [110]. These mods are either selected from among the users who are most actively involved in the community and who are invested in its success [157, 133], or self-appointed, depending on the platform. Those who become moderators due to their high level of activity usually have a better understanding of the values and expectations of the communities [157].

The nature of the role of volunteer mods can be social and communicative in user-governed communities [103]. Mods in user-governed communities play various fluid roles to shape the communities [130]; they collaborate with other mods [111] and apply moderation tools to curate content [85], and help the community leader to manage user engagement [158]. Mods also suffer from emotional tolls like lack of appreciation from the community administrator [157] and have to handle the emotional labor [44]. They might even experience impairment of psychological well-being [140] due to facing harmful content and doing the dirty work for a long time [124].

Many existing moderation systems are relying on either algorithms or human moderators that lack transparency. The algorithmic content moderation at scale suffers from opacity without explanation after content removal [59, 62]. Current work considers commercial content moderators as the *"hidden labor"* behind the scene [124], and their work is hard to be seen by the end-users [115]. Although the combination of algorithms and commercial moderators can curtail harmful content, the current moderation system on social media platforms can cause some frustration due to its black-boxed nature; for example, content removal without any explanation, appeals processes that seem to go nowhere, and minimal opportunities for users to interact directly with the administrators [115]. The challenges of the current moderation systems of social media provide an opportunity for a new moderation system that can educate and engage users at the same time [115, 86].

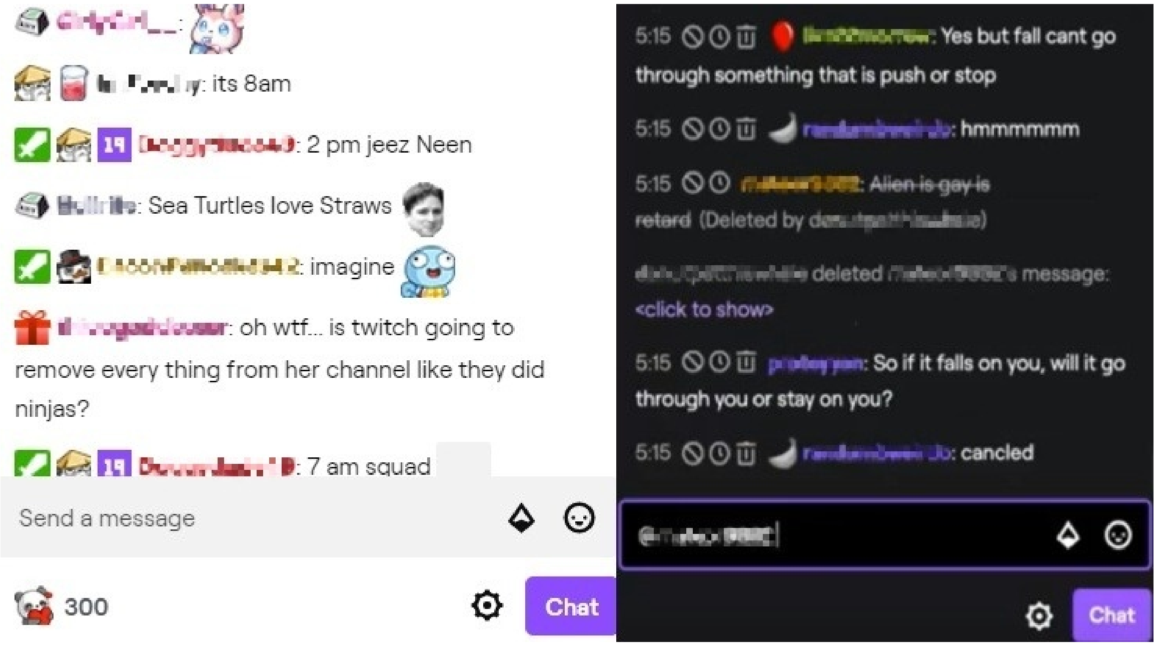
2.2 Moderation in Live Streaming Community and Twitch

As a unique social medium with high-fidelity computer graphics and video and low-fidelity text-based communication [68], live streaming is a rapidly growing industry. Twitch has become a global leading live steaming platform, starting from gaming content and expanding into a range of all imaginable content categories. In early 2020, it had more than 3 million active monthly creators and over 15 million average daily streamers [1]. It is estimated to surpass 47 million US users by the end of 2023 [2]. On Twitch, users can create their profiles under the profile settings, such as updating profile picture (displayed as a head image), adding profile banner (displayed as the background on the top of their homepages), changing username (username updates can be performed once every 60 days), and adding bio information (displayed as "About" if other users check their profile). When joining a chatroom, users can click a viewer's username to see the viewer's basic information in the channel. A further click of the username will forward the user to the viewer's homepage. When a viewer comes to the chat and starts typing, the chat rules will pop out. The viewer has to click "OK" to acknowledge the rules and to start chatting. The viewer can mention anyone in the public chat using "@" or can start a private conversation via the Whisper function under the user's profile. Twitch offers various badges to viewers to represent their status and indicate their contribution to the communities and micro-communities. Volunteer moderators have a special badge, a small icon containing a white sword with a green background. Figure 2.1 shows the interface of Twitch chatroom from both the moderator's and viewer's perspective.

To handle the user-generated content, Twitch employs a multi-layered moderation system, including both automated moderation tools and human mods, although it continues to change its structure. At a broad level, the company has employees

[1] `https://www.twitch.tv/p/press-center/`. Retrieved in March, 2021

[2] `https://www.emarketer.com/newsroom/index.php/twitch-on-pace-to-surpass-40-million-viewers-by-2021/`. Retrieved in March, 2022



[redacted]:

[redacted]: its 8am

[redacted]: 2 pm jeez Neen

[redacted]: Sea Turtles love Straws

[redacted]: imagine

[redacted]: oh wtf... is twitch going to remove every thing from her channel like they did ninjas?

[redacted]: 7 am squad

5:15 [redacted]: Yes but fall cant go through something that is push or stop

5:15 [redacted]: hmmmmmm

5:15 [redacted]: ~~Alien is gay is retard~~ (Deleted by [redacted])

[redacted] deleted [redacted]'s message: <click to show>

5:15 [redacted]: So if it falls on you, will it go through you or stay on you?

5:15 [redacted]: cancled

@[redacted]

Figure 2.1 Twitch chatroom interface from viewers' view (Left, colorful usernames with different badges to indicate status) and mods' view (Right, shortcuts of "Ban" "Timeout" "Delete" are visible next to the usernames).

who are well-trained people and mostly handle inappropriate broadcasting content that has been reported by users with common criteria for the entire community [149]. At a micro level, Twitch users form micro-communities [157] around streamers, and streamers appoint volunteer mods who are active community members to handle other users and messages in the chat with specific criteria. Since each micro-community operates under different criteria, users may behave variously across different micro-communities. Also, streamers and mods can choose to activate/deactivate a moderation tool called AutoMod that uses algorithms to filter abusive messages. Twitch also has an open-access API for the integration of thirty-party moderation tools. Mods have to track a high volume of fast-moving messages, identify the negative ones, and take action within a limited time because of the nearly-synchronous conversation in the chat [133, 157]. This poses unique challenges because it means they have very limited time to make decisions about what moderation actions they will take. The interactive social medium context with unique challenges of moderation motivated me to focus on volunteer mods and their moderation of viewers and messages in the chat.

The moderation tool on Twitch could effectively discourage spam, and specific types of negative behaviors [131]. However, the quality and functionality of bots still pose some social and practical challenges [104, 33]. Streamers also employ the help of volunteer moderators. The volunteer moderators on Twitch are appointed by the streamer and help the streamer manage the chat content. The moderators have to track a high volume of fast-moving messages, identify the negative comments, and take actions within a limited time. It is still not immediately clear how they moderate in live-streaming communities.

2.3 Conclusion

While much research on moderation has focused on the commercial moderators, asynchronous online communities, and the optimization of algorithm and the limitations of moderation tools, limited work has explored the volunteer moderators in synchronous online communities. This dissertation centers around the moderators and investigated how the moderators manage the triangular relationship among general users (viewers) and the administrator (streamer) in live streaming communities.

CHAPTER 3

MODERATION VISIBILITY: MAPPING THE STRATEGIES OF VOLUNTEER MODERATORS IN LIVE STREAMING MICRO COMMUNITIES

3.1 Introduction

Online communities provide the opportunity for millions of users to express themselves and exchange information. Freedom of speech leads to complicated challenges for these online spaces, such as hate speech and harassment. Prior literature has discussed the management of negative content from various perspectives, such as moderation techniques [152], algorithms [9], level of discourse [57], commercial labor [124], users [115, 87], policy [58], law [95], and so forth, but it is still challenging to effectively moderate these contents as the communities evolve. Live streaming, as a unique social medium with high-fidelity computer graphics and video and low fidelity text-based communication [68], is a rapidly growing industry, and also suffers from the toxic textual content. In this study, we extend previous research by focusing on the volunteer moderators' moderation practices in live-streaming communities.

Recent work of volunteer moderators and moderation mainly focuses on user-governed platforms such as Wikipedia [28, 89], and Reddit [49, 26, 85]. Twitch, as a user-moderated, live-streaming community, is similar in some governance aspects to other online communities such as Reddit, which is a self-reliant community [85], and Facebook Groups, which provides multiparty interactions [132]. However, the interactivity of live streaming makes it different from other social platforms in mainly three aspects: 1) the large volume of messages generated and posted in a short time, 2) the flow speed of these messages in the chat, and 3) the limited time for the moderator to remedy harmful situations. These unique affordances may exacerbate moderation challenges.

This study contributes a moderator-centered perspective to the growing body of literature on volunteer moderation, considering how moderators develop and apply these strategies in live streaming communities, where broadcasters showing their face have heightened vulnerability and as real-time interaction between broadcasters and viewers make harassment difficult to avoid and handle. Thus, we asked:

- **RQ:** What is the workflow of volunteer moderators in live streaming communities? Specifically, what are the strategies and how are they connected?

Through 21 semi-structured interviews with Twitch moderators, this work has mainly twofold contributions: 1) We highlight the visible activities that volunteer moderators perform during the moderation process, which has been previously described as activities that usually happen behind the scene and lack transparency; 2) We develop a diagram to show the workflow of moderation with an emphasis on the communicative components in 'live' moderation systems. We discuss how the interactivity of live stream facilitates the moderation visibility and how the synchronicity enhances the graduated moderation and amplifies the violator's voice in the workflow. Given the growing interest in using algorithmic methods to detect negativity [102], automate moderation [23], and build moderation tools [15, 131], these results provide further insight into the work of volunteer human moderators, offering potential directions into future research on the socio-technical interaction that takes place in live streaming communities as well as the design of these spaces.

3.2 Methods

3.2.1 Participant Recruitment

The project and interview protocol were reviewed and approved by the Institutional Review Board (IRB). We recruited volunteer moderators from Twitch in four ways. First, we used the official Twitter account of our lab to post a recruitment message, and at the same time, searched for moderators with search terms such as "Twitch mod" and "moderator on Twitch." If someone was interested, they could send us messages through Direct Message (a message feature of Twitter), or if we found someone, a recruitment message would be sent through Direct Message. We obtained 10 moderators through Twitter. Second, private Twitch accounts were used to reach out to four moderators by directly messaging active moderators in random channels through Whisper (a message feature of Twitch). Third, two moderators who were acquaintances or friends of acquaintances of the researchers were recruited. Last, we reached out to five moderators through the recommendation of streamers that were interviewed for another project. Each of the 21 moderators received a $20 Amazon gift card for their voluntary participation.

3.2.2 Interview

Most interviews were conducted through Discord (a VoIP communication application) with a length between 40 and 60 minutes. During the interview, we first asked general questions about moderation experience such as *"Who are you a mod for?"* and *"How long have you been a mod?"* Then we asked main questions related to our research questions such as *"How do you know how to mod?"" Do you have any prior experience?" "How do you decide what is appropriate or not?"* and *"How do you deal with toxicity and harassment?"* with the following questions like *"How do you decide when to ban, versus time out or ignore?"*. In the end, we asked them about anything that we did not mention, and they would like to share. We then closed the interview with a brief demographic indication (age, race, and gender). The beginning and end parts of the interview protocol are partially summarized in Table 3.1.

In order to have a big picture of moderation strategies and their relationship, we used thematic analysis [13] to code answers into concepts and group the relevant concepts into themes. After completing the semi-structured interviews and transcriptions, we first pasted all interview questions and corresponding answers into a spreadsheet, where all researchers went through the content of each transcript and became familiar with their content. To obtain a clear picture of themes, we grouped all the interview questions and related answers and perceptively put them under the two research questions. Second, an open coding approach was used iteratively; each researcher coded a group of interview questions and presented codes to each other in regular face-to-face calibration meetings, followed by a group discussion to clarify the consistency and accuracy. For example, the high-level category *"live explanation"* contained subcategories such as *"offering help and providing suggestions", "asking the viewer to leave",* and *"warning with prohibition"* with more detailed codes such as *"argument", "Whisper explaining", "criteria for explaining", "method of explaining ",* and *"purge or Whisper".* Then, two researchers iteratively coded all the interview questions as related to each research question independently. Finally, three researchers discussed the themes and structures and mapped them out on the whiteboard.

3.2.3 Participant Demographics

Table 6.1 lists the main demographic characteristics of our participants. Most participants were male (71.5 %), followed by female (19%) and transgender (9.5%). The average age was 29, ranging from 18 to 45. The average moderation experience was two and a half years, ranging from one to five years. The number of channels they moderated ranged from one to eighty; however, most moderators moderated less than five channels (71%).The most active among participants had a channel list that contained 80 channels. Most are moderating gaming channels, and the viewership varies from hundreds to thousands.

Table 3.1 Moderator Demographics and Moderation Activity

ID	No. of channels	Experience (yrs)	Age	Gender	Weekly (hrs)	No. of viewers	Channel type
P01	2	2-2.5	23	Male	21-84	10,000-60,000	Gaming
P02	1	2	-	Transgender	6	-	Board games
P03	6 or 7	5	31	Male	10	-	-
P04	80	4	24	Male	20	-	-
P05	30	3	21	Male	-	5-300	Gaming
P06	2	-	43	Male	Depends	few viewers	Gaming, products reviewing
P07	2	1	33	Female	20	70-130	Gaming and creative
P08	1	2	18	Male	60-70	10-100,000s	Gaming
P09	A couple	-	-	Male	35-42	2,000-30,000	-
P10	1	1.5	37	Female	3	-	Board games
P11	2	1	20	Male	21-28	5,000	Gaming
P12	1	1	21	Male	-	-	Gaming
P13	60	2.5	41	Male	21-28	500-600	Music and creative
P14	2 or 3	1	29	Male	12-16	-	-
P15	44	2	19	Male	2-3	700	Gaming
P16	20	2	40	Female	12	50-6,000	Gaming
P17	4	3-4	40	Male	4-12	200-7,000	Gaming
P18	3	4	-	Male	8-10	few viewers	Gaming
P19	5	5	27	Female	36-70	150-300	Gaming
P20	1	1	45	Transgender	16-24	100+	Gaming
P21	4	2	35	Male	30	100-500	Gaming

3.3 Results

Moderators applied a series of strategies to manage the content. We organized these strategies based on when they happen in relation to the bad act (Figure 3.1). The rectangular boxes represent a strategy. The straight lines represent a relation; the text on the straight line describes how they are related. The ovals represent an event. The diamonds represent when a decision needs to be made. The arrows represent a causal relation with the arrow pointing to the result, and the text on the arrow line represents the decision choice.

Following a time sequence, we presented the results from a proactive and reactive perspective with details such as why they used it, how they applied it, and in what situation they would use it to gain a comprehensive understanding of the moderation strategies in the live streaming community.

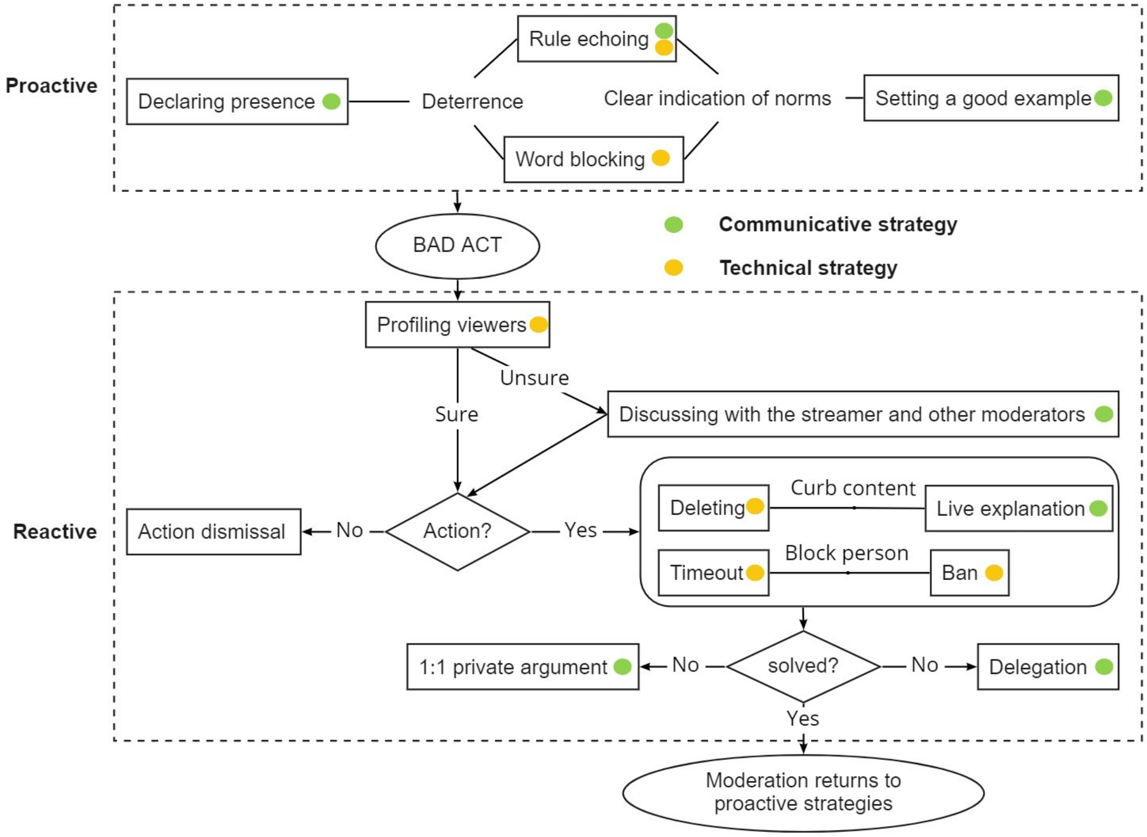


Figure 3.1 Moderation strategies for before and after a bad act happens. Lines indicate relationships, arrows indicate sequence.

3.3.1 Proactive Strategies

Proactive strategies were ones that moderators engaged in before a viewer engages in a bad act and are represented in the top half of Figure 3.1, including 1) declaring presence, 2) rule echoing, 3) word blocking, and 4) setting a good example. In this section, we described the sequence and the interactions between the elements of the diagram, which are important to understand. We emphasized that moderators' work was complex but not arbitrary. The process began with monitoring without any intervention. If moderators felt that, possibly, the chatroom could potentially go wrong, they would intervene and say something to make moderators' presence in chat visible, which could deter the potential violators (declaring presence). At the same time, moderators could keep posting the rules and guidelines manually or through the bot in the chat to remind the newcomers (rule echoing). They would also activate the Twitch AutoMod to filter obvious toxic words (word blocking). If necessary, they interacted with viewers to set a good example so that other viewers could mirror their behaviors (setting a good example). Of importance, we found that setting a good example, rule echoing, and word blocking attempted to indicate norms while declaring presence, word blocking, and rule echoing attempted to deter potential violations.

Declaring Presence Declaring presence, as a method of deterring negativity before it happened, worked as an approach of gently reminding viewers that someone who had unique privileges to enforce the rules was monitoring the chat. Declaring their presence and showing viewers that they were active by only typing a word (moderators have a special sword symbol that supersedes their Twitch identifier) would curb and deter unwanted behaviors. P07 gave us an example:

> If there was no active mod in there, people do try to push the lurk. They do say things that are inappropriate. Um, but when they see that there is even just one active mod, even if I just typed 'lol', they would see that there is a mod, that sort of cover for the trolls.

This was a communicative strategy. Moderators showing their active status in the chat by simply replying or greeting viewers deterred potential norm violators. Unlike that of asynchronous communities, the "live" element of live streaming communities indicated that the moderator was watching on-site and that any following cross-border behaviors from violators could render punitive actions.

Rule Echoing The moderator had to actively and verbally inform viewers on a regular basis because even though rules were often displayed before someone had to type, they only automatically popped out once. Streamers usually had different rules for their channels. Some were obvious, such as no sexism, no harassment, no racism, and no profanity; others might involve prohibiting self-advertising and backseat gaming (which is spoiling the game for the streamer and other viewers). Therefore, posting rules in the channel was a way of proactively communicating these guidelines with the expectation that if the viewers saw them, they should follow them. P05 thought that, since the rules were posted, then they are clearly communicated, and expected the users to *"simply follow the rules."* Yet, even if guidelines were posted on the channel, that did not mean that all users would read them. Newcomers often accidentally acted nonnormatively because they either did not know the rules or lacked experience [92]. Some moderators set up a bot that would be able to re-iterate the rules so that they would not have to type it out every time. For example, the command "!rules" would display the channel's guidelines. Using command or bot setting to post rules is both communicative and technical strategies visible to and for the public, clearly showing which behaviors are approved or disapproved.

Word Blocking Word blocking was achieved by the Twitch AutoMod that moderators could choose to activate to do some moderation tasks. AutoMod uses algorithms to hold inappropriate messages for moderators to review or prevent certain words from going into the chat. There are five levels (0 to 4) of moderation settings responding to moderation categories. Moderators could choose the moderation level and also update the terms under each level of the blocked terms list. A group of moderators reported that they liked the features of AutoMod because it could simply flag suspicious messages and reduce the workload to some extent.

If the messages were automatically filtered, only the moderator could see them until the messages were approved to the public chat so that other viewers would not be influenced. P18 expressed his appreciation for this feature:

> By far my favorite feature of AutoMod is whenever people send a message it automatically doesn't go to the chat. It [AutoMod] pretends the message doesn't exist, it turns it into a none and done deal where no one saw it, no one is reacting, there's no drama—it's gone.

This was a technical strategy. The setting and application of AutoMod happened behind the scene, and the moderation process was invisible to the public. Applying moderation tools to block words is a common strategy that has been broadly discussed in online communities (e.g., [132, 133, 85]).

Setting a Good Example Prior work has suggested that users want to fit in by doing what other community members tend to do (descriptive norms), and other community members' behaviors may be stronger indicators of acceptable ones than any explicit guidelines [92]. Similarly, we found that moderators reported being chatty, friendly, and *"answering questions"* (P19) as a way of keeping users positively engaged and hoping that viewers would imitate their behaviors. The moderator imperceptibly guided the viewers to follow the rules through this method by showing what is the appropriate language and style in the chat, resonating with Seering et al.'s work [133]. P08 said, *"They kind of look up to me, kind of follow my lead."* Similarly, P05 said,*"In moderation, people look at you for what to do, how to act, and all that. So you have to always be talking, be chatting, be helpful to people, and especially off-stream you have to be that same personality."* According to P05, setting a good example was a communicative strategy involving more engagement and visibility in the public chatroom, showing a good personality as a community member and shaping the micro-community's value. Users imitating good behaviors supported a more enjoyable chat and reduced instances of banning.

While these were proactive strategies, we noted that these strategies could also be triggered by the reactive strategies discussed in the next section. For example, rule echoing could happen from a preventive perspective, but the moderators could also post rules after they ban or timeout the violators. In addition, word blocking could be updated after the moderators observed the lexical variations of toxic words.

3.3.2 Reactive Strategies

We identified nine reactive strategies as shown in the lower section of Figure 3.1. The novelty of our findings resided in the interaction and sequence of strategies. The process began when moderators observed bad actions that violated the rules. To avoid over-reactions and maintain the community, moderators would seek to understand viewers' behaviors by reviewing chat history or applying third-party tools to track viewers' chat messages (profiling viewers). If they understand the characteristics of these viewers, but they were unsure about the punishment they should give, they would ask other moderators or the streamer for help (discussing with the streamer and other moderators). If they were sure what they should do after profiling or after the discussion with other moderators and the streamer, they would decide to either dismiss the actions and ignore these messages (action dismissal) or take a series of actions to either curb the content (deleting or live explanation) or block the violators (timeout or ban). Sometimes, certain viewers were not satisfied with the punishment and would like to argue with the moderator privately (1:1 private argument). They could also keep harassing the stream with multiple accounts so that moderators had to delegate and ask the viewers to report the violator to the platform (delegation). Till then, the moderation process was completed and they returned to using proactive strategies.

Next we first introduced how moderators profiled viewers for decision making. Then we discussed other strategies with relevant quotes to explain each strategy such as why they would dismiss actions and ignore these messages, what the standards for blocking people and curb content were, and how they interacted with violators.

Profiling Viewers The purpose of profiling was to avert mistakenly blocking a person or curbing content because suppressing expression would hinder the growth of the community to some extent. Profiling could be very quick (several seconds) or sustain a very long time (varying from minutes to hours). It played a larger role in some situations than others. Moderators learned about viewers by either observing viewers' actions for several hours on a daily or weekly basis or reviewing the chat history and the specific viewer's history. Reviewing chat history was usually achieved through technical assistance difficult to obtain from the platform. Moderators had to use third-party tools that are allowed by Twitch to assist the profiling process. These third-party tools could provide more customization than AutoMod and allow moderators to track viewers' behaviors. P18 described a tool developed by his friend: *"His most useful tool by far is what he calls a log viewer, which pretty much lets me pull logs from anytime a user has talked in a channel as long as it's been logged."*

Especially when moderators had difficulty in deciding whether to take any action, checking the log would help them make better decisions. P5 explained, *"Whenever I see a new name in chat, I'll click them and see how long they've been on Twitch. If it's a day one account, I'm immediately skeptic and I watch them like a hawk."* This information also helped moderators identify whether it was a repeated violator and decide whether it should be timed out or banned.

This was a technical strategy involving bot setting and operation to collect information. Prior work notes that moderators in user-governed communities apply various tools, including chat logs and post histories [133], but did not specify the purpose of these tools. We found that the account information and message history provided a background of the users that predicted their online behaviors. The information was helpful to the moderation action decision-making process and improved moderation accuracy.

Discussing With the Streamer and Other Moderators Occasionally, a moderator did not know how to handle the situation and had to discuss the issue with the streamer or other moderators to finally *"mutually agree"* on how to deal with it, because they did not want to *"over-moderate."* P01 explained,*"Like sometimes if we're not sure what to do [with] a person, we have a Skype chat and then we'll ask how we should deal with this person. Then we mutually agree on what to do with the person."* Similarly, P20 said,

> If there are questionable situations, we'll have discussions among the moderators or with the streamers about what to do. In niche cases where we don't know about this, we have a discussion about it on Discord or in private message about what guidelines we want to have.

Most of the time, they would directly discuss with other moderators first. Unless the situation was very serious, they would have to ask the streamers to make the final decision. P11 said,

> I don't personally talk to the streamers. It's more kind of like a general knowledge thing if they tell you something like 'you don't have to ban this guy' or 'can you ban this guy?' 'can you time this guy out?', whatever. It's more of that kind of interaction. We don't personally have meetings with the streamers unless it's something super serious like a sponsorship or anything like that.

Prior work has suggested that in user-governed online communities, moderators often apply an open discussion for changing rules in communities with less structured hierarchies, and the head moderators can arbitrarily make final decisions without asking for feedback in communities with a clear hierarchy [133]. We found that in live-streaming communities, it was the streamer, not the head moderator or other moderators, making final decisions.

Action Dismissal After moderators had a basic understanding of the violators, they decided to ignore violations in some cases when they knew the viewers' persona, perceived viewers' intentions (to receive attention from others), or just decided to distance themselves from the situation.

Some viewers would always behave in a certain and expected pattern. In some situations, the moderator or the streamer had already classified these viewers' personas. With the streamer's approval, they decided to disregard these behaviors by doing nothing, even though those viewers violated the rules. P02 explained,

> There's this guy. He likes to be toxic but then they're saying that's his personality online, like an online persona. It's just weird to me. It's just something I have to put up with.....Then I told [the streamer] about it and then [he] told me yeah that's just his personality. I said it's weird to me but okay.

Some viewers broke the rules in order to get attention from others. Moderators elaborated that the best way to deal with these attention seekers was to ignore them and their inputs in the chat. *"Sometimes they're just looking for attention and sometimes you just ignore them; they just go away,"* said P19. The reason was that *"any further attention paid to them, it's just gonna feed them more. They're gonna continue trying to do it,"* said P17. Sometimes, the negative content caused heated discussion and increased the interaction in the chat. If the misbehavior was not very serious and the moderators thought it did not cross the line, they decided not to take action. P05 gave an example: *"Usually, if it's a really terrible troll I'll ignore them, then let them humiliate themselves and let chat have fun with it."*

We found that moderators used "let it go" as a strategy to distance themselves from the violator. P04 said, *"The easiest thing is if you have trolls trying to get through your skin you kind of let it go and laugh it off."* Specifically, some moderators took short breaks to leave the screen and let these negative contents go with the chat flow instead of taking any further action. P07 shared her experience:*"I'm just going to go on a quick cup of tea. I'm having five minutes to myself and then went back."*

Action dismissal or non-response to violators is an atypical response to anti-normative behaviors. According to Figure 3.1, this is neither a technical nor communicative strategy, only one that involves cognitive processing. To a certain extent, high interactivity in the live chat results in the messages being transitory so that even though moderators did not take any action, the negative messages would disappear as more messages emerge. This strategy appropriately reduced information overload and emotional labor of moderators, but we did not know how the ignored content would affect other viewers. In order to minimize the negative impact of trolls, it has to be a widely followed norm of recognizing and ignoring them [92]. However, the challenge is that ignoring requires considerable self-control not to respond to offensive provocation, especially for new community members [73]. Thus, moderators may also need to educate viewers to identify and ignore these trolls, not just isolating themselves.

Live Explanation Recent work shows that Twitch users perceive educating as an effective strategy to get rid of toxicity [17]. Moderators explained the rules to viewers through live explanation and education. Unlike simply deleting with a warning, live explanation involved more engagement and offered help and suggestions to violators. The purpose of doing this was to build the community and curb the inappropriate content in the public chat without any punitive actions. Moderators often applied this strategy when they saw public argument among viewers or the chat topic became sensitive and was considered inappropriate for the public.

The public chat area is not a suitable place for arguments because it is mainly used for common topics that everyone can get involved with as well as interact with the streamer. An argument between two viewers could disturb the chat experience for other viewers as well. P08 said, *"If two people are arguing in the chat, I always [tell] them to take it to their DMs or Whispers or whatever to handle it there because the chat is not the place to do that."* P08's explanation is consistent with prior work that has indicated that moving conflicts to special locations where the normal rules of behavior do not apply will be met with less resistance from users [92]. The Whisper function of Twitch offers a private space for one-on-one interaction.

Though some topics did not violate the rules, they were considered not appropriate in the chat because they were too personal or sensitive and could bring down the vibe and potentially cause negative impacts. Moderators dealt with those viewers by either providing resources they could utilize to help the viewers or politely asking them to leave, in an effort to protect the remaining viewers. P16 stated that she would remind these viewers to cease their actions:

> There are some people who are negative because they're depressed. They come out like with their guns blazing and everything, and you tell them to knock it off, and they kind of back down pretty quickly. And you know, just speaking with them privately, you suggest that they get some help. I have phone numbers bookmarked for if people need someone to talk to, that sort of thing.

The direct explanation between moderators and viewers could also rectify the misbehavior before it went beyond control and finally got the user banned.

Moderators would gently remind the potential violators to remedy minor offenses. P20 said,

> If they say something that they may not understand right. For example, sometimes people will walk in and will say something like, 'oh hey you're really pretty' and that's not an acceptable behavior so usually we will not ban them, we will say to them, 'hey that's objectifying and that's not an appropriate comment, it's not respectful to comment on the looks of a streamer so don't do that again'.

P07 reported a similar tolerance: *"If they are less offensive and just being cheeky or maybe pushing a little bit, you send them a whisper and say, look, you know, calm down a little bit."* *"Usually the user will listen and apologize for it,"* P12 noted.

The educating and suggesting in both the public and private chat was a communicative strategy, either maintaining the chat atmosphere or rectifying lightweight violation. Prior work has discussed the black-box nature of the current moderation system and the lack of an educational system [115]. Live streaming communities integrate the explicable and educational components into the moderation process. The synchronous nature of the live chat provides an opportunity for immediate feedback of the moderator's conduct to the viewer and also the viewer's performance to the moderator, making the education and explanation process possibly efficient. Our finding supplements Jhaver et al.'s work on Reddit that explanation of removal is under-utilized in moderation practices [86] and educating users with helpful feedback improves user attitude of fairness and intention to post in the future [84].

Deleting Content, Timeout, Ban These strategies were commonly applied as moderation activities. We found that in "live" communities, deleting happened when the viewers did not read the rules of the chat and incidentally said something inappropriate. Even if moderators decided to remove these messages sometimes they did not ban the violator with an expectation that they would not perform the same behavior. P07 said, *“Those that just fail to understand what they’re saying, it’s either rude or something, we’ll purge what they said.”* Sometimes, deleting was followed by an explanation or warning, resonating Jiang’s work of moderation in live voice communities [88].

Also, warning messages came in various forms of intensity. Some moderators used a gentle tone, reminding the viewers that such behaviors were not allowed, such as, *“Hey, we don’t use that kind of language,”* said P12. Other moderators stated using severe sentences, cautioning users of the punishment awaiting them, should they proceed with their unacceptable actions. P03 said, *“You get that warning like ‘Hey FYI, don’t do this again otherwise you’ll get ten-minute time out and then, you know, a third strike and you’re banned’.”*

A temporary ban, usually referred to as a “timeout,” was a less severe solution for misbehavior compared with a permanent ban. Moderators reported having people in the chat who were mostly positive and respectful but might misbehave and cross the line. Temporarily banning the viewer who broke the rule sent a message to the viewer and the rest of the community that such behavior was not welcome. P05 stated, *“If you can tell someone has the intent of being a good community member, but they’re a little overbearing, then that’s a timeout.”*

Several moderators deliberated that spamming emotes and text in the chat would get a timeout, which is different from Facebook that sends warning messages to the users and Twitter that investigates account activities, removes from search, or terminates the account [59]. P11 said, *“If someone is spamming the same message a couple of times, I will probably just time them off for ten minutes or so.”*

A permanent ban meant that the users would never be allowed into the stream again. Not only was it a severe punishment for the user, but moderators also used this command sparingly because it affected the overall viewership. However, many moderators mentioned that they had *"zero tolerance"* toward obvious and severe issues such as racism and sexism and would ban these behaviors, similar to prior work [88, 133].

In addition, since live-streaming communities are streamer-centric, anything potentially harming the streamers and their benefits reserves severe punishment. Any personal attack toward streamers' appearances was also a permanent ban. Inappropriate comments such as the *"streamer's bad"* or the *"streamer's ugly"*, resulted in a permanent restriction on the viewer's ability to watch the stream (P08). P01 similarly reported, *"They might just like attack the players, whether physical appearance or lie how they play. Obviously, if it's physical appearance, then I have to purge them or ban them."*

One participant specifically mentioned that self-advertising of other streams deserves a permanent ban. In Twitch, many streams are similar in the content they provide, especially gaming streams. Thus, there is usually a lot of competition and a tendency to promote one's stream on other channels. P19 said, *"You actually do a permanent ban if they're advertising their stream in a chat. I don't have any type of tolerance or patience for that."* According to P19, the competition among different micro-communities escalates the moderation sanction. The content of self-advertising is not as severe as racism, sexism, or personal attack, but allowing it impairs the community, thus moderators have no *"tolerance or patience."*

Sometimes moderators had different tolerance levels toward the same violation. For example, dealing with trolls in the chat was viewed differently by moderators. P21 said, *"You time someone out if they are troll. They will just leave because they don't want to wait ten minutes again and again."* But other moderators would permanently ban the same act. P06 said, *"But if someone is clearly just there to troll or just be* *a Jerk. Those people, there's nothing you can do with them, and there's no saving them. You just have to send them on their way."*

Generally, the deleting, timeout, and ban are technical strategies invisible to viewers and fitting the "graduated sanctions" [119], beginning with persuasion and light sanctions and proceeding to more forceful actions [92]. As parts of reactive strategies, the multi-level sanctions based on the severity of misbehaviors increases the legitimacy and thus the effectiveness of sanctions [92].

1:1 Private Argument Viewers have the opportunity to argue with moderators through the private message; these conversations often happen during the stream. Sometimes viewers attempted to start arguments with moderators regarding the grey area between what was and was not allowed in a private chat. These arguments usually took place after a punitive action due to a user's misconduct in the chat. Viewers would argue that they should not be banned or timed-out through Whisper, and the moderators would argue the reason and deal with it on site. For example, P03 stated:

> [The viewer is] being rude and being deliberately rude. Like the rules say don't be an XXX, and that's exactly what he was being... he kept bugging me, he's like 'well that doesn't really explain why you did what you did' and I said, 'Quite frankly, I'm here to do my job. I'm not here to be your friend.' I've said that before, and that's the ultimate thing.

According to P03, the private chat allows the violator to express his opinion even after he was publicly banned. This process increased the perception of procedural justice, and the legitimacy is enhanced by providing users opportunities to argue their cases with the moderator [92]. The moderator was forced to perform in real-time in the private chat, which requires improvising. This was a communicative strategy, increasing the visibility of moderators in front of violators in the private chat. The nuanced difference between live explanation and private argument was that live explanation focused on the education and explanation in both public and private chat while 1:1 private argument focused on the debate between moderators and violators in the private chat only.

Delegation The moderators also encouraged other viewers to report violators because moderators could only process and deal with a limited amount of negative messages and problematic viewers even with the assistance of moderation tools. In certain situations, when the problematic viewer intentionally tried to disrupt the channel and created many accounts to harass the streamer or moderators, the moderator suffered from limited cognition and failed to address all issues in the chat. The information overload was difficult to handle in these situations. A smart approach to follow was to utilize the power of the crowds. Some moderators would ask viewers for support and do a "live crowdsourcing" to moderate chat comments. P13 said,

> Maybe try to encourage viewers to go ahead and report this user, so hopefully, they get an IP ban. Those are only in really extreme cases when somebody won't go away, because Twitch is bad at that. If a viewer wants to create 50 accounts and harass someone privately, it's very hard to prevent that.

This strategy was a communicative strategy seeking the public to engage, similar to moderation techniques encouraging users to flag suspicious content and report to the platform on Facebook [35] and relying on users as witnesses to collect evidence of rule-breakers in voice-based communities [88]. The reason behind this act was that volunteer moderators wished that the platform administrators (commercial moderators) could intervene since they might have more power to ban the IP of the violator.

3.4 Discussion

This work mapped out the moderation strategies applied during the moderation process, contributing to the growing body of discussion about volunteer moderators and moderation in HCI and CSCW. We want to clarify that our main contribution is not the novelty of the strategies, but rather, it is the flow of how these strategies happen and the decision-making processes of moderators in the live context.

The interactivity of live streaming meant that moderators have to combine proactive and reactive strategies that engage both technical and communicative solutions, suggesting that moderators had to deal with harmful content in front of viewers on-site, explain and educate violators publicly or privately, and discuss with other moderators and the streamer behind the scene. These activities were accompanied by the challenge that because of the real-time nature, large volumes of content lead to information overload and only allow limited time for decision making and multi-task handling during the event. In the following section, we discuss how the unique affordances of live streaming increase the visibility of content moderation.

3.4.1 Interactivity Facilitates the Visible and Performative Activities of Moderation

Different from commercial content moderation that mostly happens behind the scenes [124] and lacks transparency [115], moderation relying heavily on volunteers increases the visible and performative activities. Among the 13 strategies in Figure 3.1, six involve technical, and seven involve communicative strategies. Technical strategies usually operate behind the screen and are less visible to viewers, while communicative strategies are mostly in the public chatroom visible to everyone or in private chat only visible to a specific violator. Only one strategy 'rule echoing' was found to fit both categories, where it is both a communicative and a technical strategy. Many communicative strategies applied at both proactive and reactive level can be achieved because live streaming provides an interactive and immersive experience for user engagement [67].

Moderators are usually the glorified viewers who are actively engaging in and influencing the channels [157]. During the streaming event, they still watch the stream as the viewers do, but with an eye on the chatroom. At the proactive level, the moderator sometimes needs to interact with viewers in the public chatroom by answering questions or joking around. This kind of performance happens in parallel to the performance of the streamer. In this sense, moderators are *the* viewers and *interacting with* other viewers. When they saw potential harmful actions, their roles would change to law enforcers who discretely dealt with the situation without disturbing the chat. They would either declare presence or post rules to deter these behaviors. Thus, the publicly visible activities involved different roles as moderators had to toggle between being the face of socialization/ community role model and justice enforcer. At the reactive level, the moderators have to explain and educate violators and delegate moderation tasks to viewers in the public chat or argue with violators in the private chat, indicating that the visibility of moderation increases moderators' vulnerability to negativity and violators [142]. How to balance moderation visibility and moderator protection should be further investigated.

Generally, volunteer moderators in the interactive context perform much visible communication in the public chat and private chat than commercial moderators do. The role (moderator, viewer) dynamic and visibility of volunteer moderators highlight the importance of affordances of live streaming when considering their roles and transparency and appear to be more prominent in the live streaming context in comparison to other social media platforms.

3.4.2 Synchronicity Enhances the Graduated Moderation and Amplifies the Violator's Voice

We echoed some moderation strategies broadly applied in online spaces such as content removal and banning the end-user [139, 87]. However, most of these common strategies are working separately. In most cases, one action is the end of moderation, such as content removal without rational explanation [115, 84] or directly banning the community [27].

According to the diagram in Figure 3.1, we systemically connected these moderation strategies and displayed them in a sequential flow to clearly show how moderation works in this new type of community. Kiesler et al. [92] applies the "graduated sanctions" concept in online community settings and suggested that the lowest level of sanctions is a private message explaining the violation, where sanctions escalate after repeated or more severe misbehavior. This concept can only partially explain connections of reactive strategies but not proactive ones. Thus, "graduated moderation" seems to be more appropriate to include the proactive strategies in the workflow. The simultaneity and ephemerality of live streaming not only require instant attention and immediate moderation (e.g., one minute delay in moderation response could lead to a chaotic chat environment) but also make graduated moderation more effective than that on asynchronous communities because the moderators are always actively watching during the streaming event. The graduated moderation starting from proactive strategies instead of simply excluding violators shows the much effort moderators put to minimize the actions that could potentially alienate community members. Thus, graduated moderation increases the legitimacy and the effectiveness of moderation in the live context.

Moderation work in live-streaming communities empowers viewers to actively engage in the chatroom because the synchronicity brings everyone in the channel actively online all the time. The asynchronicity of most online communities limits the interaction of users and moderators and causes difficulty or delay for users to acquire feedback and guidance in time. Users lack the motivation to actively seek feedback unless moderators actively post explanations or contact the users. The delayed feedback discourages meaningful social engagement and relationship building. Prior work also points out that end-users develop their own folk theories configuring what is appropriate [39] because of the lack of explanation after content removal in online spaces [84]. In live-streaming communities, end-users can play larger roles than in asynchronous communities during the moderation process. For example, once a message was deleted, the viewer could ask the active moderators on-site for the reason or argue with the moderator that it was unfair. Thus, their voices can be heard by moderators in the dynamic interaction process and their valuable feedback may potentially contribute to the moderation process. Prior work has suggested that community influence on rule making increases compliance with the rules [92]. Therefore, community influence in live streaming plays a larger role on rule making than that in asynchronous communities, thus resulting in possibly higher compliance with the rules.

As new platforms emerge with novel technology, they may also take on property above currently unique to live streaming and consider how the moderation workflow works. For example, moderators in voice-based communities, such as Discord, secretly record voice for evidence and take extreme actions of excluding such as muting and banning [88]; instead of taking reactive strategies, moderators can combine some proactive strategies such as echoing the rules with declaring presence. The moderators can orally explain the rules or even have a recorded rule explanation to broadcast now and then in the voice channel. Though the diagram of moderation strategies is complex, it clearly shows the mental model of moderators. We can explicitly see where the decision making takes place and which strategy has been explored broadly or needs more attention.

3.4.3 Design Implications

We propose that designers and developers should consider advanced technical tools to facilitate the profiling process. Current tools can only provide limited information about the viewers through the log function. Future tools should be able to provide more performance data of viewer's activity such as how long they have been online; how frequently these viewers communicate in the public chatroom and argue with moderators in the private chat; and the ability of tagging messages of the viewers' characteristics such as funny, talkative, elegant, well-behaved, toxic, and trolling, similar to the tagging mechanism on Twitter [78]. These data can help moderators increase the understanding of viewers and save time to make more accurate decisions during the moderation process.

An algorithm or system to identify the violators' type should be considered for moderators to make action dismissal decisions. We know that if the moderator knows the viewer's characteristics and intentions, they take no further action. For example, developers can design a classification system that can: (1) identify these problematic viewers based on text messages or chat history, (2) classify these viewers into specific categories such as attention seekers and a viewer saying bad words with good intention, and (3) annotate these messages and viewers and notify the moderators. This kind of system would reduce the monitoring effort and automatically catch violators when a large volume of messages pour into the chat, especially when moderators are handling a particular viewer and cannot keep an eye on the chat.

Communication is critical for effective moderation in live steaming communities, but the communication tools in the system were sub-par. We found that not all moderators would use the private messages function for discussion; they also used external tools such as Discord and Skype. Usually, a streaming channel has multiple moderators to ensure that at least one or two moderators are online when the streamer is. The problem, which is an opportunity for improving the design, is how the outcome of discussions between active moderators and the streamer can be documented so that other inactive moderators can be well-informed without wasting time checking the whole conservation history across different tools, which is simply an attempt to reinvent the wheel. It will be helpful if there is a system or feature that can automatically summarize the discussion in bullet points or highlights and save it as a document that can be shared with all moderators. Zhang and Cranshaw have developed a prototype system for Slack to automatically summarize chat conversation and share it with group members [161]. It is promising to bring such design to live streaming communities.

A documenting system would facilitate communication between not only moderators and streamers but also moderators and viewers. Explaining the rules through live interaction involves a lot of typing and interaction with viewers, which is time-consuming, and due to the limited cognitive abilities of the human brain, moderators might potentially overlook other negative content in the chat, causing a deterioration in the moderation job. If there is a bot or feature that can document these explanations in the system, and easily call out a specific explanation when necessary, we speculate that moderation efficiency would be highly improved by just simple 'click and send' instead of repeatedly typing. For example, we categorized 'rule echoing' as a communicative and a technical strategy. Since the content is already available in a written format, re-posting the relevant rule (as opposed to posting the entire rule list) when necessary, would help streamline the moderation process and increase the chances of viewers actually reading the automatic message.

3.4.4 Limitations

There are several limitations to this study. First, our participants are volunteers, not commercial moderators. In order to generalize the findings, further research can focus on commercial moderators in live streaming and compare the differences. Because the governance structure of each social media is different, we think it is inappropriate to claim that the user-moderated model in Twitch is similar to commercial moderation found in platforms like Facebook. Our findings may apply to other communities that have user-governance with simultaneity such as Discord, live-streaming communities, or live VR communities, but not all online communities. Also, even though our sample shows diverse moderation experience, our sample has more male than female and transgender participants. We are not sure if gender is a factor that influences moderation.

3.5 Conclusion

We identified the flow of decision-making that takes place during the moderation process. These practices of volunteer moderation bear similarities but also distinct differences compared with other user-governed communities. The interactivity and synchronicity of live streaming reveal the visible and performative work of volunteer moderation. This work reminds us to think about moderation from another perspective. Instead of considering moderation as blocking content or violators with the assistance of technical agencies, we may also want to take social dynamics into the moderation process and highlight the significance of communicative strategies performed by the human moderator at both the proactive and reactive level. The affordances of live streaming also allow graduated moderation and amplify violators' voices in the moderation process, showing moderators' great effort to increase legitimacy and maintain community members.

CHAPTER 4

CATEGORIZING LIVE STREAMING MODERATION TOOLS: AN ANALYSIS OF TWITCH

4.1 Introduction

Technical interventions can, to some extent, reduce the human moderation load, especially in large and fast moving chats [3]. Many online communities, such as Reddit and Twitch, apply bots (software robots) to assist the mods in doing moderation practice [133]. Current research about using bots for content moderation mainly focuses on asynchronous communities such as Reddit [56, 104] and Wikipedia [33, 114], with limited research about bots for moderation on Twitch [132]. Better understanding of the moderation tools that mods use every day would help improve the current tool design, reduce the working load of mods, and further benefit the community. The goal of this research is to analyze the features of moderation tools on Twitch into categories that could be generalizable to all other moderation tools and to provide some implications for future tool design. I used the same data collected in Chapter 3, 21 moderators on Twitch with diverse experience to answer the following two questions:

- **RQ1:** What kind of moderation tools do Twitch mods use in live streaming?
- **RQ2:** What do mods expect from moderation tools in the future?

4.2 Results

4.2.1 Moderation Tools

Based on moderation tools that they used, moderators could generally be divided into heavy technology users or light technology users. Most of them were heavy users, and if they used bots, they usually used more than one and the combination varied. Some

Table 4.1 Moderation Tool Categories and Functions

Chat Control	*Viewer Control*
Chat movement control: P1, P9	One click and purge: P1
Multi moderating: P18	Ban or timeout: P2, P13
	Pause without timeout: P8
	Log view: P5, P18
Content Control	*Settings Control*
Flag and alert message: P1, P5, P20	Filter words: P2, P6, P18
	Customization: P2, P5, P8, P18

were light users and stated that they did not like bots and that the bots often caused more trouble so that they mainly moderated manually and only used the basic bot embedded in the system.

The most popular bots or extensions that our participants used were: Nightbot (38%), Twitch AutoMod (33%), Better Twitch TV (BTTV) (33%), Moobot (19%), individually developed bot (19%), and FrankerFaceZ (FFZ) (10%). Among these, only the Twitch AutoMod was built into the Twitch system; others were third-party plugins or extensions. (Although AutoMod is in the Twitch system, users can choose not to activate it if they do not want to use it). Interestingly, some participants mentioned they were using tools that they or their friends developed. Then, the authors categorized these tools regarding their features. Based on participants' description, four categories and nine examples of the features that fall into those categories are summarized in Table 4.1.

Chat Control Some moderation features were associated with control of chat, a place where viewers could comment on streamers and communicate with each other.

The chat interface is side by side to the live stream (on PC it is on the right, on mobile devices the chat is on the right or beneath the video, depending on whether the device is held vertically or horizontally) and happens simultaneously.

Because of the live interaction on Twitch, all the new messages sent by viewers would be automatically displayed at the bottom of the chat, making it challenging to go back and check chat history if new messages were constantly appearing. The inconvenience of going back caused difficulty for some mods. *"When you go on Twitch, and you try to delete a message, and you scroll up, if somebody sends a new message it automatically goes to the new message,"* P1 explained. Some extensions could help them control the speed of the chat movement. P1 added: *"There is a tool that makes it when you scroll up it does not go back down."* In big channels with lots of viewers, the chat moved so quickly that they could not catch negative comments—for situations like this there was a feature that could make the chatroom still. P9 said, *"I have an extension where if I hover over the chat with my mouse, it just stops the chat, so I can properly click on someone's name and moderate."*

Content Control Flagging and alerting "bad" messages was a feature mainly integrated into Twitch AutoMod. P5 explained this feature:

> I... turn on AutoMod, which is Twitch's automation thing because all that does is flag messages as pending. So, if a message is deemed inappropriate by your channel, it'll flag it and then put it in chat for the moderators. They can say approve or deny.

In addition to flagging and alerting messages, the system could also automatically filter certain words. Moderators or streamers could set and put filter words in bots so that these words or the variants of these words typed by viewers could not be displayed in the chatroom. *"You can put specific words into it that just don't go through,"* said P2. P18 expressed his appreciation for this feature:

> By far my favorite feature of AutoMod is whenever people send a message, it automatically doesn't go to the chat. What I really enjoy about automod is that it pretends [the message] doesn't exist, it turns it into a none and done a deal where no one saw it; no one is reacting; there's no drama—it's gone.

Viewer Control There were many features in controlling viewers' behaviors. "One click and purge" allowed moderators to easily and conveniently delete the offensive message and "time out" viewers from the chatroom simultaneously. P1 said: *"It is easier to purge people because it is just one click and you purge them or ban them whereas on Twitch you would have to actually like type it out with like purge or timeout or ban. So, it allows you to do things more conveniently."* The ability to do something with "one-click" indicated the efficiency of using the moderation tool.

If someone said something inappropriate, some words that have been considered too toxic or offensive by streamers or moderators, the ban or timeout rule would apply. This feature was mainly implemented through extensions. *"I use BTTV, and that gives some nice things to make it easier to time out and ban,"* P13 said. Nightbot also had a similar function, filtering words first and then timing out the person. P2 said:

> Nightbot tries to make sure if someone says "faXXot" it just does not appear on Twitch. It just... that person will end up timed out. It automatically times out the person from being able to talk for a specific number of seconds. I believe it's 60; I'm not sure.

Pause without timeout was a little different from and less severe than a ban or timeout. Instead of timing out a person for a specific period, a pause would slow down the speed of messages that one could send. P8 said:

> Instead of choosing to permanently ban somebody or time them out for 10 mins in chat, much time you will see a mod purge somebody, which is just literally to time them out for one second, and I have this setup... in my settings that I have a button to set people's name that I can automatically purge them without actually time out like slash timeout.

A pause without timeout worked as a light warning. The messages had no problem, but someone might want to get attention and, instead of typing a sentence that might be overwhelmed by others' messages, might type quickly word by word to take up multiple lines. Then the whole chatroom would be occupied by the messages. These messages would annoy other viewers and dilute community experience.

Log view allowed the moderators to check a specific viewer's log. By doing so, they could see the chat history of the viewer. *"His most useful tool by far is what he calls a log viewer, which pretty much lets me pull logs from anytime a user has talked in a channel as long as it's been logged,"* said P18. Especially when some viewers were discussing lightly harmful topics, but the moderators had difficulty in deciding whether to give a warning, a timeout, or a ban. Checking logs would help moderators to make better decisions. P5 explained:

> You can look up people, see how long they've been following. We can see previous chat messages; you can see all tons of information about them. So, whenever I see a new name in chat, I'll click them and see how long they've been on Twitch. If it's a day one account, I'm immediately skeptic and I watch them like a hawk. Otherwise, I just let them chat.

Settings Control Many moderators discussed that customization of settings based on their needs made moderation more efficient. *"It is more efficient. You can customize the tool whichever way you want, and it's just a lot better for people,"* said P8. Some bots provided the option to customize timeout, for example. *"A common plugin for*

Twitch, you can add custom timeout buttons for different tasks," said P5. Similarly, P8 said,*"For external tools sometimes I use custom IRC clients if I want to run like a custom bot to look for a specific keyword to time out."* Some bots allowed customized settings to track details of chat activities. P2 said:

> When I created my own (setting), it's like, it's very detailed. It tells you everything that happened, even while you're not in the chat. Something that will happen a week ago, it'll be like this is what went down.

Even though current bots provided a certain level of customization, from our interviews, some moderators were not very satisfied with the performance of customization. More options for current features such as timeout settings could be considered higher-level customization as well.*"The Twitch tool, it is mostly being able to do it one second, 10 seconds, or say one second, one hour, or 10 hours or whatever. That's pretty much it. Like it does need to be more in-depth than that,"* said P19. These deeply customized features would meet moderators' diverse needs, reduce their workload, and accelerate the moderation process.

Through the analysis of current moderation tools, nine features were highlighted, and four categories were identified. However, are these all they wanted? Are there any other features they expected? The following research question asked about moderators' needs.

4.2.2 The Desired New Features

Our second research question was about what mods desired in the future. The question specifically asked the mods in the interview: *"If someone could design a moderation tool or bot for you, what would you want it to do?"* Since not all moderators have used all existing tools in the market, some wanted features that already exist and were covered in the previous section. Thus, in this section, only new features not mentioned above will be discussed. Six features were identified: grouping chat by languages, having a pop-out window to hold messages, chat speed control, a set of buttons with pre-written/ pre-messaged content, viewer activity tracking, and all-in-one. Ironically, some of these features were already available with existing bots or extensions, but the participants were unfamiliar with it.

Grouping Chat By Languages This feature was relevant to the content control category but different from any features mentioned above. There were many viewers from different countries, speaking different languages, but watching the same streaming event. Not all viewers would type and communicate in a single language. However, if the moderators only understood one language, it would be difficult for them to moderate when the content of different language mixed. Moderators might be distracted and have to pick out messages that they could read and understand, even though different moderators were assigned to handle different languages. Therefore, they wished to have a function to group different languages for different moderators. Doing so would improve moderation efficiency. Moderators also wanted translation abilities to help out with chat in different languages. P1 said:

> Like, because Gears of War is so big in Mexico, and it's just a lot of people who speak Spanish are in the chat. Sometimes it gets overwhelming to the point where the American, or the people who speak English only. They might not have anything to do in the chat because we just can't understand what's being said. So maybe a feature on Twitch or Mixer that automatically [translates] Spanish, or any language in general, to English would be cool and helpful for us so that the people who only speak English, or not only speak English but predominantly speak English, could help along. It also helps the moderators who speak Spanish because now they have so much more work to do because it's not equally divided among us. So, they have a heavier workload.

Pop Out a Window to Hold Messages This feature could be under chat movement control category but was different from the features mentioned above. A pop out window would hold the message that the moderator wanted but would not change the chat flow. The participant said Twitch once had this feature, but after the update to the latest version, it was gone. Now it was hard to hold messages. P7 said:

> I think popout would be very good. If you could make it, so a bot could make a pop-out window so that when you click on something it would hold it. Now you don't get to pop out where you can inspect what the person is saying. If I could get a bot to bring that sort of thing back. Because if you can go back and look over the sort of things somebody saying if they're just swearing and it's a one-off assessing something inappropriate, it's a one-off. It's not such an issue, but if I can go back and see that this person has insulted X, Y, and Z, I think I said something inappropriate to someone and its little things, then you know, you've got to keep mind.

Chat Speed Control This feature was also relevant to chat control but different from other features mentioned earlier. P8 said that the messages moved so quickly and were hard to catch up. However, he only hoped a new feature to slow down the speed so that he could not click and moderate by mistake. Something might look like an audio player, and there are options such as slow down, keeping normal, speed up. He explained:

> The chat moves quickly, so you want to slow it down... If I want to timeout someone and someone posts, the chat is going to go up like one line, so I can ban someone else by mistake.

A Set of Buttons With Pre-written/Pre-message Content This feature could be under settings control but was different from the customization features mentioned above. Again, some bots already have this feature, but the participants were unaware of them. Participants mentioned that they would need commands with pre-written information so that they could reply more quickly than just typing the same message again and again. *"I would just press a button, and it would instantly reply with something, that I had pre-messaged or pre-written,"* said P11. With this setting, moderators could work more efficiently. P15 described his expectation and said:

> I would probably have it be like go all around so it would probably have stuff I'd take inspiration from night bot you know having commands with info, so having that ready... obviously, it'll be quicker than us since it is a bot and not a person.

Viewer Activity Tracking This feature could be under the viewer control category but is different from the log view. In the log view, moderators wanted to check one specific viewer's chat history and make better judgments based on the viewer's current performance. Viewer activity tracking was about the general behavioral summary of a group of viewers. For example, what percentage of them are super active? How many of them are lurking? How many new viewers joined in or left last week?*"I would want something that would track everyone else. I want some vocal data about regular people, get notices if people do not show up. I can notice if people suddenly get depressed, maybe that,"* said P21. Many moderators expressed their care about their viewers during the interview and considered some of the viewers as friends and had a good relationship with viewers. By owning this feature, moderators and streamers could have a better understanding of viewers' activities. Therefore, they could improve their service and maintain a better relationship with viewers and doing so would be beneficial to the community as well.

All in One This was not a novel idea, but moderators wanted something that integrates all the features of moderation tools in the current market into one. P4 moderated for several big channels and had to use five bots to assist the moderation process because currently, no one tool could meet his requirements. He explained:

> I think it'd be cool to have an all in one moderation bot where you can type in a name and give it like a Twitch whisper or something else, so you could pull it quicker than you could from going through a website or chat logs in a program.

4.3 Discussion

The first research question identified four different perspectives taking the synchronous nature of live streaming into consideration, preliminarily providing a guideline for further bot development in this domain, and the second research question supplements the four categories identified in the previous one. Similar to Seering et al.'s findings [132], viewers control involves a certain level of multiparty interaction between moderators and viewers. Future design can explore how to facilitate the interaction and at the same time improve moderation efficiency. Our results also show that the moderation tools in synchronous online communities are different from those in asynchronous online communities such as Wikipedia and Reddit. For example, chat control involves real-time content management, and mods have to deal with information overload and to make decisions immediately, suggesting that mods in live streaming communities are undertaking a different type of time-sensitive psychological pressure than those in other communities.

The categorization of moderation tools enables us to think about features in a more systematic fashion, not only in identifying the different types of problems that exist, but also where more work needs to be done. According to the analysis of features of current moderation tools and features that moderators expected, the authors have several suggestions for the design of the platform as well as suggestions of new features.

4.3.1 Design Opportunities

Specifically, for Twitch, the leading live streaming platform, the main features of its AutoMod are mostly under the content control category, which means that features under the other three categories are opportunities for future development. Twitch allows third-party extensions, thus opening up opportunities for a myriad of different moderation tools. However, it is still difficult for beginners to choose which tools to use. The beginners might have to add so many extensions to test one by one and then only keep the better ones. If Twitch can add a function to categorize tools by their features, it would be helpful for moderators, especially beginners, to search for the tools that they need.

Some moderators expressed the desire for features that already exist, indicating that searching for these third-party tools is inefficient or that there is a lack of information about where to find extensions or bots that are less well known. Future research may want to look into how moderators discover these tools, but the fact that people do not know about tools that already exist means there are more opportunities for centralized repositories of these tools and education about how to use them.

Technology updates so quickly. Some unavailable features during the initial research planning and execution of this study are now available on Twitch. For example, some interviewees mentioned that when they scrolled up the chat, it would automatically go back down. However, now when scrolled up, the messages will stay where they are stopped. Twitch also has a “Popout” window to hold chat and to run separately. Moderators can keep both the chatroom embedded in a streaming webpage and the “Popout” window open and can use the chatroom to track general behaviors of viewers and the “Popouts” to deal with suspicious viewers. The evidence further exemplifies the importance of understanding the function of these features from a higher perspective than the feature themselves. The identified categories are not time-sensitive.

4.3.2 Suggested New Functions

Based on some of the frustrations and problems that moderators discussed, the authors suggest a couple of ideas for new features that could be applied to any live streaming platform.

Highlighting the content moderators want to track: a language setting button that allows moderators to choose what kind of language would be highlighted on their screen that will help them focus on what they can handle and increase working efficiency. For example, a Chinese moderator would only want to moderate Chinese content in the chat and click the button to show Chinese messages only. All the Chinese would be highlighted, and other language content would turn gray or shadowed so that they could concentrate on the moderation of Chinese content.

Instead of checking viewer’s log (which would mean that during that time the moderator would be ignoring the whole chat to moderate problematic viewers), a setting similar to the language setting could be applied as well. If the moderators thought a specific viewer was suspicious, they could be able to click on the viewer’s name, and all the messages from this viewer would be highlighted (e.g., in red color) in the following message flow. One click and starting to track the subsequent behavior would amplify their capability of moderating. However, the prerequisite is a setting that can slow down or speed up the chat movement so that moderators can accurately identify the problematic viewers.

Content and rule category setting: this feature is inspired by multi moderation and applied for different channels and content, but it could also apply to any single channel. It means that one bot can have a setting that contains many different rules and streaming content categories that accommodate the norms and guidelines of different channels. From our interviews, the rules for teen channels did not apply to adult channels. In adult channels, adult jokes were permitted but might be inappropriate for teen channels. Hence, settings that can choose a content category first and then apply a rule for that specific category would improve moderation accuracy and avoid embarrassing situations and negative impressions.

4.3.3 Limitations and Future Work

Our sample is grounded in just one online platform—Twitch. Further research can take other live streaming platforms into account and confirm that transferability of the results across similar live streaming services. It is also important to note that very few social media platforms have a governance structure in place that allows for third-party moderation tools. That said, it would be interesting to know what kind of moderation tools are being used by companies that do not have third-party tools. Besides, the authors randomly recruited participants on Twitch but finally obtained more males than females and also included some transgender participants. The biased gender toward males may have an impact on the results. Since gender difference is beyond the scope of this study, further research may explore the tool or feature preference among these genders. Lastly, many other potential perspectives on the themes of moderation tools can be triggered. For example, future research can explore how to facilitate communication among viewers and mods in the viewer control theme, and chat control theme might need further investigation to understand better how to reduce information overload of mods in the live streaming community.

4.4 Conclusion

Through the interviews with a diverse sample of moderators on Twitch, the authors used a grounded theory approach and identified four high-level uses of moderation tools that provide a method of conceptual categorization that can potentially apply to any live streaming platforms. Through the summarization of mods' expectation of tools in the future, several functions that can fulfill mods' needs are identified and support the four perspectives. Since multiparty-based chatbots are underexplored, this research provided many insights into bot development in the live streaming community and raised issues related to social interaction among moderators and viewers, community norm evolution, and technical development of moderation tools. Live streaming is still growing very fast, and content moderation for it is still a challenging issue. No existing bot is perfect to meet the moderator's needs, indicating that there is a potential market and opportunities for related bot development.

CHAPTER 5

WHAT ARE EFFECTIVE STRATEGIES OF HANDLING HARASSMENT ON TWITCH? USERS' PERSPECTIVES

5.1 Introduction

The live streaming platform Twitch applies both technical intervention and human moderators [133, 157], but is unique in that there are more opportunities to self-govern compared to social media such as Twitter or Facebook. Moreover, the communities are centered around the streamer, who has some control over what people are permitted to say. In this study, we ask users of Twitch who should take responsibility for handling harassment and which strategies they think are effective:

- **RQ1:** Who should be responsible for deciding how to enforce what is appropriate?
- **RQ2:** What are effective strategies in getting rid of harassment behavior?

5.2 Methods

The survey data was collected during TwitchCon, an annual convention for Twitch enthusiasts that is hosted by Twitch. Six researchers walked around the convention and asked attendees (mainly people standing in line for something) to fill out a paper survey. Participants were given a small, custom pin that we designed for completing the survey. The survey included questions about their favorite streamer (not a part of this study) and about content moderation on Twitch (items development based on the pilot interview and brainstorm). Results from the paper surveys were then put into Survey Gizmo for digital archiving and subsequent analysis.

5.3 Results

The sample (N= 375) was mostly male (64.2%), 23.3% female, and two people who identified as non-binary. Age (*M*= 26.05, *SD*= 6.56) was between 12 and 52 years. Of the 80% of participants who reported race, most were White (44.4%), followed by Latino/Hispanic (13.1%), Asian (12.7%), Black (4%), Pacific Islander (3.7%), and other (2.1%). 59% said that they were a streamer.

To answer the first research question, we asked in the survey, "How important are the roles of the following entities in terms of deciding how to enforce what is appropriate to say in chat? Please rate from 1 (not important at all) to 5 (very important)." The frequency table (see Figure 5.1) displays users' responses. Participants thought that the streamer should be the most responsible with the highest average score (*M*= 4.67, *SD*= .85), followed by the moderator(s) (*M*= 4.23, *SD*= 1.04), the company (Twitch)(*M*= 3.47, *SD*= 1.48), and the viewers(*M*= 3.13, *SD*= 1.39). The differences between the results were statistically significant (Table 5.1). Also, independent t-tests showed no difference in results between the streamers and non-streamers.

Table 5.1 Difference Test

Pairs	T-value	P-value
Streamer vs Moderator(s)	7.11	.00
Moderators vs Company	8.51	.00
Company vs Viewers	3.46	.00

To answer the second research question, we asked "How effective do you think are the following strategies in terms of getting rid of toxicity? Please rate from 1 (not effective at all) to 5 (very effective)" in the survey gave participants a list of strategies based on our earlier qualitative work. We conducted a Principal Components Analysis with Varimax rotation method and eigenvalue greater than one. The exploratory

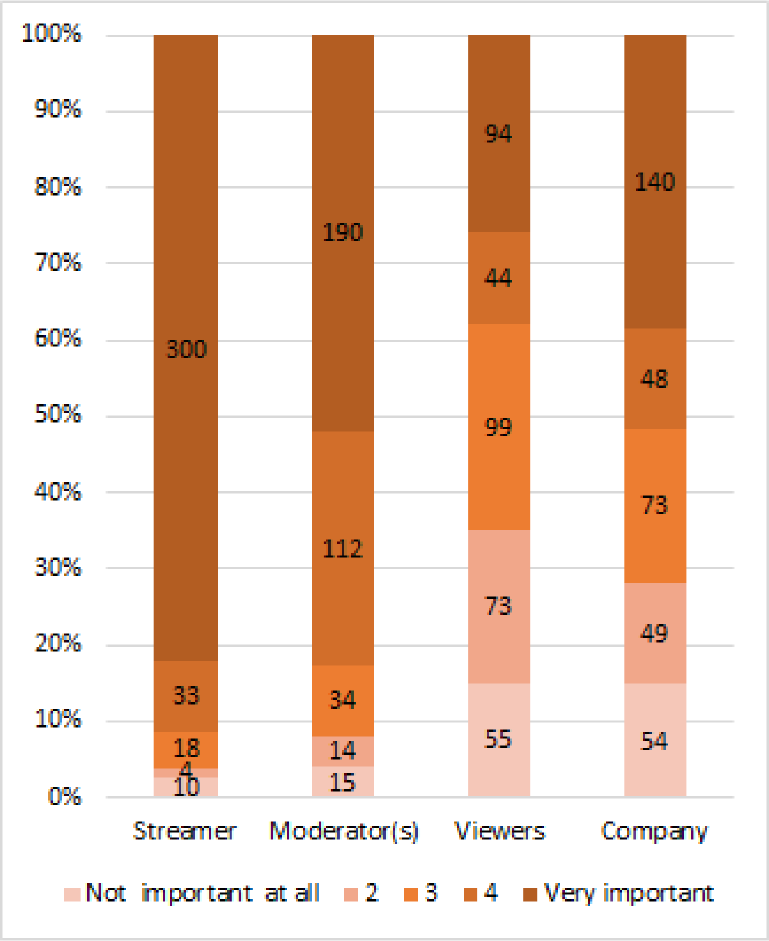


Figure 5.1 Who should be responsible for deciding how to enforce what is appropriate?

factor analysis revealed five factors with a total explained variance of 68% (see Table 5.2). According to the description of items, We named these five variables: Educating (*M*= 3.10, *SD*= 1.15, α= .82), Sympathizing (*M*= 2.02, *SD*= .94, α= .76), Shaming (*M*= 1.68, *SD*= .91, α= .67), Humor (*M*= 2.62, *SD*= 1.26, α= .74), and Blocking (*M*= 4.01, *SD*= 1.03, α= .62). Educating refers to telling or explaining to the violator how to act appropriately. Sympathizing refers to caring about the violator and trying to help. Shaming refers to responding to the violator with the same toxicity. Humor refers to laughing off the toxic comment. Blocking refers to banning the toxic person from speaking either temporarily or permanently.

Table 5.2 Exploratory Factor Analysis of Effective Strategies

Themes with Items	Loadings				
Educating					
Explaining to the toxic person how to act properly	**.88**	.16	.08	.02	.15
Educating the toxic person on the rules of the stream	**.84**	.10	-.08	.00	.17
Telling the toxic person what they are doing is wrong	**.83**	.11	.06	.05	.10
Asking the toxic person if they are feeling okay	**.52**	.44	.06	.18	.00
Sympathizing					
Trying to have a discussion with the toxic person	.14	**.78**	-.07	.06	.01
Sympathizing with the toxic person	.08	**.77**	.08	.02	-.04
Asking the toxic person why they are toxic	.14	**.76**	.20	.14	.07
Extending pity to the toxic pity	.14	**.58**	.38	.14	.01
Shaming					
Saying rude things to the toxic person	.10	.08	**.87**	.04	-.04
Shaming the toxic person	.03	.05	**.72**	.15	.14
Being toxic back to them	-.08	.16	**.71**	.18	.01
Humor					
Responding to toxicity with humor	.08	.10	.10	**.90**	.01
Treating toxic statements as a joke	.04	.16	.29	**.82**	.01
Blocking					
Banning the toxic person from the stream	.11	-.06	.10	-.04	**.86**
Timing out the toxic person so they can't chat for a certain period of time	.22	.08	.01	.05	**.80**

We also asked participants to write in any strategies that were not listed above. The open-ended question revealed several themes to supplement the factor analysis results (Table 5.3). Many participants suggested to "simply ignore them" (M, 30), and this strategy is effective because toxic people just want attention. A participant explained: "Ignoring even if they are not banned or timed out. If they do not get a reaction, they will go somewhere where they will." Not only would they ignore the toxicity, but also would be "telling the viewers to ignore them" (F,23).

Table 5.3 Other Strategies to Combat Toxicity

Category	Code Count
Ignore	50
Encouraging positivity	13
Tolerance before ban	11
Making rules clear	7
Having good mods	5
A combo of options listed in Table 2	3
Bot intervention	2
Asking the community to help curb it	1

Participants also suggested to "promote positivity" (F, 24) such as "teaching the toxic person how to be positive" (M, 28) because "positivity breeds positivity" (F, 37). The streamer or mods should encourage "positive conversation and foster a healthy community" (F, 28) and "have everyone involved in the community engage in a positive and friendly way" (F, 20) when things do not go their way.

Many people were willing to give opportunities to first-time violators. For example, "Just be kind, give them a chance, continue with a ban if it continues" (F, 29), "Extend a second chance to first time offenders, but after that, a ban is in order" (F, 25), and "Once they have been reported three times, impose a 30-day ban"

Table 5.4 Quotes of Other Strategies

Making rules clear
Persistent and consistent applying the rules (M, 37).
Making sure your community is all on the same page of what is acceptable in your chat so they can help set the correct tone and support the chat while you are streaming (F, 32).
Having good mods
Having good moderators that understand your wants in getting rid of toxicity in chat along with a supportive community (F, 24).
A combo of options listed in Table 2
We usually time them out for 10 mins, tell the person what they did wrong then give them a chance to come back and stay (M, 23).
Bot intervention
Posting help links with bot commands (M, 39).

(F, 35). Similarly, “Track who bans by profile, not just in the channel, after three bans on different channels, either ban the profile or make a toxic emotes, although that is a form of shame (sad face)” (M, 50). Other quotes are displayed in Table 5.4.

5.4 Discussion

Twitch users thought that the streamer should be the most responsible entity to enforce the rule in the chat instead of the company; it would be interesting to see how this compares to users of social media like Facebook and Twitter. One possible explanation is that the live streaming community has a decentralized governing structure, and the users generate and moderate content autonomously.

Among the five strategies identified in the factor analysis, blocking and educating were the most effective strategies, and the other three (humor, sympathizing, and shaming) were perceived as less effective with the average score under three. Interestingly, we found ignoring was a popular strategy that was unprompted but mentioned by many users. It might be caused by the attribute of real-time interaction in the live streaming community and the fact that conversations are somewhat ephemeral. Without any action, the toxic messages in the chat will soon disappear as more comments emerge. Moderation on Facebook and Twitter often happened behind the scenes so that it is easy to block but difficult to educate the problematic viewers. In the live streaming community, the live interaction in the chat allowed moderators to block while educating at the same time. Design to facilitate educating and blocking or to help moderators to balance ignoring and actual educating and blocking should be considered. The attendees from TwitchCon were experienced users with an in-depth understanding of moderation, gaining insights into answering our research questions, but the limitation was that they would not represent the average Twitch users.

5.5 Conclusion

In this study, we asked users about who should be responsible for deciding how to enforce rules on Twitch and found that they held the streamer to be most responsible. We also conducted a factor analysis to identify five strategies (educating, sympathizing, shaming, humor, and blocking) and the open-ended questions revealed several more strategies (ignoring, encouraging positivity, tolerance before ban, etc.).

CHAPTER 6

AFTER VIOLATION BUT BEFORE SANCTION: UNDERSTANDING VOLUNTEER MODERATORS' PROFILING PROCESSES TOWARD VIOLATORS IN LIVE STREAMING COMMUNITIES

6.1 Introduction

Since community growth and health is about not only punishing but also maintaining users by setting positive examples [131], understanding users' characteristics is a good way to avoid sanctioning users by mistake and to improve the perceived justice and fairness. Checking a user's account information, which is a good indicator of a user's characteristics (e.g., [51, 69, 148]) and a reference to the user's commonalities with others (e.g., [99, 136]), is one way to do so. As account information and activities reveal users' behaviors, profiling, which refers to the dynamic process of collecting and integrating users' information and activities to find their behavioral patterns and characteristics [96], is vital for mods to understand bad actors, a challenge highlighted by prior research [87, 84].

In line with recent HCI and CSCW research proposing to understand bad actors [90, 12], this work aims to explore how mods psychologically profile violators in live streaming communities. Due to the lack of HCI theories associated with profiling, we used criminal profiling [74] as a lens to understand moderators' mental models about the profiling process and types of violators. To achieve this goal, we first observed moderation work through mods' self-recorded videos. Using videos as probes, we then interviewed mods while watching the videos. In the interviews, we asked their reasoning to deal with violators, such as the information they are looking for and the reason for their judgment.

This work contributes to understanding mods' mental models regarding what happens after violation but before sanction. In most cases, profiling allows mods in micro-communities to understand violators' characteristics to avoid excessive punishment or, more importantly, mediate and support community members. We present their profiling process, how they collect information for profiling, and the violators' types with various moderation strategies. We discuss how the platform's affordances and design affect profiling; we also discuss how profiling can potentially grow the community through increasing justice and fairness and distinguishing bad actors. Finally, we suggest social and technical interventions that could assist in profiling in the moderation process.

6.2 Related Work

6.2.1 Criminal Profiling

Criminal profiling, also called "psychological profiling" or "offender profiling," is "an educated attempt to provide investigative agencies with specific information as to the type of individual who committed a certain crime" [54]. Similarly, Egger [47] defines criminal profiling as "an attempt to provide investigators with more information on the offender who is yet to be identified." Generally, criminal profiling is the process of gathering evidence both at the scene of a crime and from the victims and witnesses to construct a biographical sketch of the criminal [97]. Hicks and Sales [74], in their book dedicated to the development of criminal profiling, propose that crime scene evidence is the primary source of investigative information available to investigators, including physical evidence and victim information and statements, and that the offender's characteristics cause them to leave particular pieces and patterns of evidence during the crime. Through these shreds of evidence, the investigator pieces together the offender's characteristics to figure out the types of offenders.

Focusing on the roles that evidence can play in informing a timeline and narrative of the crime, Chisum and Rynearson [30] classify physical evidence into different types such as sequential (sequence of events surrounding a criminal act), directional (where something was going and coming from), location (position and orientation of people and objects surrounding the scene), and limiting (the nature and boundaries of the crime scene) evidence. The breakdown of evidence can facilitate answering “who,” “what,” “ when,” “where,” “how,” and sometimes “why” questions about the commission of the crime [31]. The evidence is usually collected by the crime scene investigation team consisting of photographers and specialists and then sent to forensic psychologists and other experts to analyze [116]. Criminal profiling shows how a group of researchers systematically collect evidence and deduct criminals’ characteristics based on the evidence and is worthwhile for police investigation because of its improvement in the scientific rigor of research (e.g., see the meta-analysis by [45, 52]).

Criminal profiling plays different roles in the criminal justice system in three phases: criminal investigation, apprehension, and prosecution [74]. In the investigation phase, profiling aims to link evidence as part of a series to identify physical and psychological characteristics of unknown offenders, to predict the pre-and post-offense behaviors that an offender might show, and to evaluate the potential escalation of certain criminal behaviors. In the apprehension phase, profiling suggests evidence collection on the search warrants or interrogation techniques eliciting a confession from an offender and predicts an offender’s behaviors on the arrest. In the prosecution phase, profiling works as providing expertise in the courtroom to demonstrate the linking of multiple offenses to one individual or to match a particular individual to the relevant crime(s) [74, p13]. In this study, we mainly focus on the investigation phase, which aims to understand violators’ characteristics or evaluate violators’ behaviors to avoid similar violations happening in the future.

6.2.2 Applying Criminal Profiling to Community Moderation and in Live Streaming Communities

Much research in HCI discusses profiling normal users online, such as how to develop different types of clustering, how to use algorithms to cluster users and develop different personas (e.g., [125, 29, 46]), and how to predict users' preferences and provide better services (e.g., [143, 113, 144, 64]). Stainbock [141] reveals the connection of general profiling using algorithms and criminal profiling and states that "data mining's computerized sifting of personal characteristics and behaviours (sometimes called 'pattern matching') is a more thorough, regular, and extensive version of criminal profiling." In these contexts, the person conducting the profiling is usually an industry professional and targets the regular user but not violators. Little research in user profiling literature focuses on collecting the moderated information to profile violators, the information that is removed and invisible to the public. Mods have access to both the normal content visible to the public and the invisible violative content, owning the advantage to see the holistic scenario to understand a user's behavior and characteristics. Though some work focuses on collecting moderated information to understand violations, no specific work applies the profiling lens to understand violators.

In community moderation, criminal profiling has been used as a lens to exemplify how Wikipedia moderation tools work as profiling agents, from observing and catching vandalistic edits to finally generating patterns using either structured decision-making or a black-box approach [38]. Additionally, some research points out the necessity for human mods to collect evidence for their decision-making during the moderation process. For example, Jiang et al. [88] found in live voice chat on Discord, mods face challenges to collect evidence of potential violators, sometimes even with the risk of violating privacy policy to secretly record voice as evidence. Kiene et al. [91] also found that the moderation tools are insufficient for organizing and retrieving information for mods to make consistent decisions toward violations and that mods seek user-developed bots to track information of community members. Research in live streaming communities shows that some mods use moderation tools to check a viewer's history [19] but are not satisfied with the features of these tools and hope to have more information about viewers and violators [15]. While this thread of research discusses the need and necessity of more evidence for moderation, they do not specify what type of evidence they need, how they collect the evidence, and consequently, how to use these shreds of evidence to evaluate potential violations and punish potential violators.

The need of understanding evidence collection in community moderation and the lack of framework in user profiling literature to understand violators indicate the potential of a new lens to build a connection between community moderation and profiling research. Additionally, much research also introduces various types of justice (e.g., social justice, retributive justice, restorative justice) from the criminal justice system to explain online harassment and moderation, and the justice-seeking process [128, 127, 10, 140, 38]. The inherent role of criminal profiling in the criminal justice system and its components (evidence collection and analysis, and the deduction of offender's characteristics) in the definition suggests that criminal profiling may serve as a good lens to understand mods' profiling process when they face potential violators at a conceptual level. In line with the definition of criminal profiling and applying it to live streaming communities, we asked the following questions:

- **RQ1:** What kind of evidence do mods collect to profile violators?
- **RQ2:** How do mods collect these types of evidence?
- **RQ3:** What are the types of violators that mods perceive?

The online environment makes the application of this lens slightly different from the offline world. First, online behaviors become part of the evidence. In online communities, harmful content is considered crime scene evidence and reflects online behavior. Most types of evidence relevant to physical evidence (e.g., blood, body drag, glass fragments) are not applicable to the online environment. Second, the offender is already identified in live streaming communities, so the profiling is not to find the offender but to evaluate whether or not the mod should punish them and to what extent the punishment should be. Instead of directly banning users, they may also look for other evidence. Third, researchers in criminal profiling rely heavily on the captured offenders' self-reported information to figure out their characteristics. In live streaming communities, mods as non-experts directly communicate with violators and can access various information.

6.3 Methods

Since the profiling process happens behind the scene, we chose the observation plus interview method to explore the research questions. The observation allowed us to see the whole moderation process, from seeing a violation to finally sanctioning the violator. The interview alongside the video allowed us to recall the moderation actions with mods and then to ask questions about their decision-making process. The school IRB approved this project, and the consent form was sent to participants before the interview through either email or Discord.

We offered two options for mods to participate. The first option (A) was to share with us a self-recorded video of the screen when they were moderating. After we reviewed the video, we scheduled the interview. Mods received a $100 Amazon gift card after the interview. Because some mods had strong privacy and safety concerns and/or felt uncomfortable with recording, we provided a second option (B) with a $50 Amazon gift card, only conducting a semi-structured interview but asking them to provide necessary examples (e.g., screenshots, video clips) during the interview.

6.3.1 Participant Recruitment and Demographics

We recruited 19 participants through three approaches. First, we reached the potential participants through the email list that we collected from Twitch Convention 2019 and received six responses. Twitch Convention is a gathering of the Twitch community hosted by Twitch to provide the opportunity for streamers, moderators, viewers, and merchandisers to meet offline. We recruited from Twitch Convention offline to increase diversity and minimize the bias of only recruiting people online. Second, one research assistant who was also a Twitch mod asked other mods in the channels he moderated to recruit five participants. Third, we used our personal Twitch accounts and browsed the recommended channels on the Twitch homepage. We first entered live channels to observe for 5-10 minutes. After we saw active mods, we asked and obtained eight mods. We had 12 male mods and seven female mods. The average age was 23. Most mods were white. The average moderation experience was three years, ranging from half a year to eight years. Most primarily moderated gaming communities. 10 mods chose option A, and nine mods chose option B. The viewership of the channel in option A varied from tens to thousands. Details are summarized in Table 6.1.

6.3.2 Video Analysis and Interview Process

We first ran a pilot study with the mod in our team. Three researchers interviewed the mod to test the flow of the interview protocol and watched the mod's moderation practices to decide the reasonable length of recorded video for the observation. The pilot video was one and a half hours long, with 105 active viewers on average in the chat. We observed ample violations and repeating moderation in the full video, even in the first hour, and thus considered one hour a reasonable length for the observation. All the participants were encouraged to share with us a one-hour-length video through Google Drive. To analyze the video, we focused on moderation related actions and

Table 6.1 Demographic and Experience of Participants

ID	Option	Viewership	Category	Experience (yrs)	Age	Race	Gender
P1	A	18-20	Gaming	4	21	Hispanic	F
P2	A	10-15	Gaming	4	19	African American	M
P3	A	70-100	Art, body painting	2.5	23	Hispanic	M
P4	B	—	Gaming	1.5	18	White	M
P5	B	—	Gaming	3	27	African American	F
P6	B	—	Gaming	3.5	34	White	F
P7	A	30-35	Rhythm & music game	0.5	18	White	M
P8	A	15-20	Gaming, video editing	4	18	White	F
P9	B	—	Gaming	1	19	White	M
P10	A	130-150	Gaming	2	18	Asian	M
P11	B	—	Gaming	3	19	White	F
P12	A	650-1400	Gaming, IRL	2	21	Asian	M
P13	B	—	Gaming, IRL, Drama	3	29	White	M
P14	B	—	Gaming, IRL	8	28	White	F
P15	A	800-1000	Gaming, IRL, eSports	6	31	White	M
P16	B	—	Gaming, IRL	3	24	Pacific Islander	M
P17	A	9000-11000	Gaming	1.5	21	White	M
P18	A	3000-4000	Gaming	1	20	White	F
P19	B	—	Gaming	5	26	Asian	M

developed a codebook for video coding (1, explain; 2, delete; 3, warning; 4, timeout; 5, ban; 6, should have moderated but not (ignored); 0, other interesting issues). An explain is the rule explanation in the chat; a delete means the message was removed in the chat; a warning means sending a warning message to the viewer in the chat; a timeout is a temporary block from minutes to hours; a ban is a permanent block, indicating the violator can not send a message in the chat anymore. Warning, delete, timeout, and ban can be achieved via bot command that is alongside the username and badges, as shown in Figure 6.1. In the coding process, we focused on these actions and excluded mods' social interaction, such as greeting newcomers and just chatting with viewers. Each video was analyzed by two researchers separately to identify the timestamps of each relevant action. Then, the two researchers discussed their results to achieve consistent timestamps.

All the interviews were conducted after the video analysis and through Discord. Before the interview, we opened the recorded video on our side and also asked mods to open the video on their side. In the interview, we first asked some general questions about their moderation experience, such as which platforms they moderate for and how long they have been moderating. Then we asked some questions about providing examples of moderation decisions they made. Later, we asked them to look at the video for each timestamp that we noted and to explain their decisions. For example, "At 35:04 (35 mins and four secs), I saw you deleted the message and banned the user. What was your rationale to make that decision?" For mods who chose option B, we skipped these questions. After questions on video analysis, we asked questions about profiling, such as what kind of information helped them moderate and what the reasons/motivations were for users to perform badly. Since option B did not share video to help us gain context, we often asked follow-up questions such as "do you have a specific example to show us?", "can you give us an example?" and "can you explain more about this?" These follow-up questions reminded them of something they recorded and saved from their end. They thus shared the content with us via Discord during the interview. In the end, we asked for demographic information. The interview protocol and process followed a consistent structure for both options, except that the video plus interview option added several questions for each timestamp, and that the interview-only option asked more follow-up questions about examples. All interviews were audio-recorded, transcribed by speech recognition software[1], and then double-checked by the researchers.

[1] `https://www.temi.com/`. Retrieved on March 14, 2022

6.3.3 Interview Analysis

We imported all transcripts into ATLAS.ti Cloud[2] for collaborative coding. First, four researchers individually went through all transcripts to have general ideas. Next, four researchers picked up a transcript with abundant content to code individually. After individual coding, four researchers had a group meeting to discuss the codes and clarify the definitions. All codes with definitions were archived. Four researchers, repeating the above steps, coded three transcripts to develop an initial codebook. By following the initial codebook, each transcript of the rest was coded by two researchers individually and discussed later to achieve consistency. During this process, any new codes were added to the initial codebook with a definition. The other two researchers then reviewed the new codes and their definitions for agreement and applied the updated codebook to code the next transcript in sequence. After finishing the coding, the authors exported codes to a spreadsheet to iteratively organize relevant codes under each research question to form subcategories and categories (see supplemental files).

6.4 Results

Before identifying the violations, mods usually monitored the chat and sometimes interacted with viewers. Sometimes, they could not define whether the messages in the chat were violations. They waited for more information to evaluate the purpose and meaning of the messages. P3 (M, 23) said, *"It's difficult sometimes to ascertain things, but as long as people aren't saying, 'Oh you look awful', or things like that, I'll usually leave, and I'll try and gather more information and see what they're going to do in the chat because more often than not, I can't predict the future, at least give them the chance to talk."* For lightweight issues, several mods reported that they tended to give people chances and *"watch and see"* (P12, M, 21). Once the mods

[2] `https://atlasti.com/cloud/`. Retrieved on March 14, 2022

confirmed the violation in the chat, the profiling process was triggered and involved evidence identification, evidence collection, and violator type formation with possible punishment.

6.4.1 Evidence Types

To answer the first research question, ‘What kind of evidence do mods collect to profile violators?’ we adopted physical evidence types [31, 30] from the criminal profiling framework and identified three types of evidence applicable to online communities. According to our observation of mods' activities, they conducted evidence collection in a very specific sequence. We followed the sequence of how mods processed information and presented this section.

Action Evidence Action evidence refers to information that reflects the online behaviors. For example, in Figure 6.2, the message "fresk is bad" was considered action evidence that reflected the violator's intention and behavior to harass the streamer and was deleted. This user specifically pronounced the streamer's name and said the streamer was *"bad."* Mods reported many different types of behaviors/acts, such as being malicious, trolling, spam, racism, and sexual perversion. These violations have been broadly discussed in prior work (e.g., [49, 73, 60]). In addition, mods also noted that disruptive behaviors such as nonsense-talking in the public chatroom broke the synchronous experience, though these messages did not break the rule. If the disruptive behaviors went far and caused trouble to other users, mods would step in and sanction these behaviors. Mods sanctioned violators differently after they turned the one-time offensive action into repeated offenses, such as P18 (F, 20): *"Sometimes they don't realize that their message is offensive, but people like that who says things impulsively. I know their intention. So I just delete it, and if they keep going, I just give a timeout."*

Ownership Evidence Ownership evidence refers to information that reflects the identity or source of the violator. It consisted of offensive usernames and throwaways accounts, username position, badges, channel status, and account status. Though a few types of the above evidence were more or less mentioned in live streaming research (e.g., badges and throwaway accounts), we considered them necessary components to represent the holistic picture of the profiling process and explain them from the profiling and moderation perspective.

Offensive Usernames and Throwaway Accounts Usernames were observable evidence that was directly and visually collected by mods. Before the violation happened, mods in most cases considered offensive usernames as heuristic indicators of the potential violation and tried to avoid their influence in the community. Offensive usernames *"indicate more that they're there to cause trouble rather than to actually participate"* (P8, F, 18). These users circumvented the rules, and the username display was too offensive to be consistent with the channel's value. For example, P12 (M, 21) shared two offensive usernames via Discord during the interview: a sexual username like *"Ice_wallo_come"* meant I swallow cum, and a sexual harassment username toward underrepresented groups like *"ray_ping_minors"* meant raping minors. Mod worked with the streamer to *"ask for them to switch over to a new account if they want to watch"* (P7, M, 18). In most cases, offensive usernames can be considered indicators of potential violators and paid special attention to these users. However, in some cases, it was context-dependent, and mods relied on other clues to figure out the purpose of users.

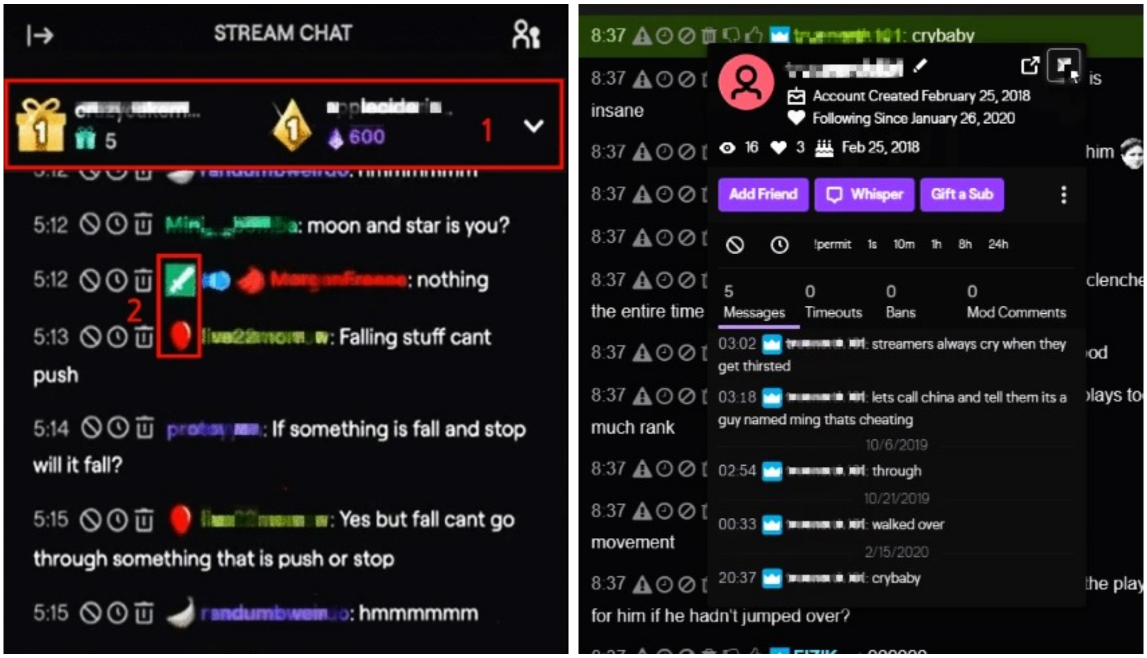
After the violation, an offensive username alongside a negative message (action evidence) provided additional information and enhanced mods' judgment on whether the user was an intentional violator. They noted that they would carefully check the user's account information. P2 (M, 19) expressed his logic: *"I typically immediately* *click a toxic name with a toxic message. That makes sense."* Mods also used usernames to identify what they perceived to be throwaway accounts. P17 (M, 21) described that accounts with *"a bunch of numbers "* were *"obviously throwaway accounts."* In order to further judge whether it was a throwaway account, P5 (F, 27) stated that she would *"go through and check their profile and see if it's like blank or anything."* If the account history was empty, there was a high chance that the suspicious account was a throwaway account. Unlike typical throwaway accounts using letters and numbers, some accounts directly harassing others by saying something negative about a specific user or the streamer could also be considered offensive usernames.

Username Position The username position on the Leaderboards [3] (a Twitch feature for the streamer to give viewers' recognition by pinning gifters' and cheerers' usernames to the top of the chat window, as shown in Figure 6.1a Box 1) also played a role. P3 (M, 23) explained that a big donation made the username appear on top of the chat for a certain amount of time, indicating support and contribution to the community. Mods recognized these usernames, had a higher tolerance when these users violated the rules, and were less likely to punish them, compared with users not on the Leaderboards.

Badges Along with a username were the badges owned by the user (As shown in Figure 6.1a Box 2). Twitch offered users different types of badges [4] specific to the channels, such as cheering chat badges (users purchasing virtual currency-bits and paying bits for special animated emotes to cheer the chat and support the streamer) and subscriber badges (users paying a monthly fee to support the streamer and owning

[3] `https://help.twitch.tv/s/article/leaderboards-guide?language=en_US`. Retrieved on March 14, 2022

[4] `https://help.twitch.tv/s/article/twitch-chat-badges-guide?language=en_US`. Retrieved on March 14, 2022



5:12 [redacted]: moon and star is you?

5:12 [redacted]: nothing

5:13 [redacted]: Falling stuff cant push

5:14 [redacted]: If something is fall and stop will it fall?

5:15 [redacted]: Yes but fall cant go through something that is push or stop

5:15 [redacted]: hmmmmmmm

8:37 [redacted]: crybaby

[redacted]

Account Created February 25, 2018

Following Since January 26, 2020

16 3 Feb 25, 2018

5 Messages 0 Timeouts 0 Bans 0 Mod Comments

03:02 [redacted]: streamers always cry when they get thirsted

03:18 [redacted]: lets call china and tell them its a guy named ming thats cheating

10/6/2019

02:54 [redacted]: through

10/21/2019

00:33 [redacted]: walked over

2/15/2020

20:37 [redacted]: crybaby

8:37 insane

8:37 the entire time

8:37 much rank

8:37 movement

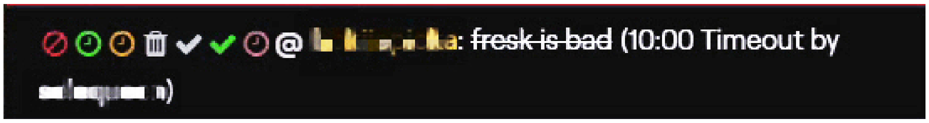
8:37 for him if he hadn't jumped over?

(a) (b)

Figure 6.1 Screenshots of the interface from a moderator's view integrated with different moderation tools. (a) Box 1: the LeaderBoard; Box 2: badges; the icons alongside each blurred usernames are three commands: ban, timeout, delete. (b) A moderator checks a user's message history: this user sent only 5 messages in the chat with 0 timeout, bans, and mod comments.

different badges by subscribing different lengths). An active user usually owned various badges, either through purchasing subscriptions to the streamer or for free (e.g., VIP badge denoted by the streamer to recognize the loyal members). Usernames with badges were less suspicious than those with no badge. P12 (M, 21) described, *" The first thing that stands out to me about a user is if they have badges or not. That's like, whether you're a subscriber or if you have Twitch Prime, or if you have anything. If you have a badge, generally speaking, it's less suspicious than just a regular account with no badge."* P12 also explained that an account with no badge would make him *"curious"* and *"click"* it.

Channel Status Channel status refers to the user information and activities in the channel (micro-community). Mods also reviewed channel status, specifically, following



⊘ ◷ ◷ 🗑 ✓ ✓ ◷ @[redacted]: ~~fresk is bad~~ (10:00 Timeout by [redacted])

Figure 6.2 A message deleted, and a user timed out by P18. The icons alongside the blurred usernames are commands without any badges, such as ban, two different timeouts, delete, etc.

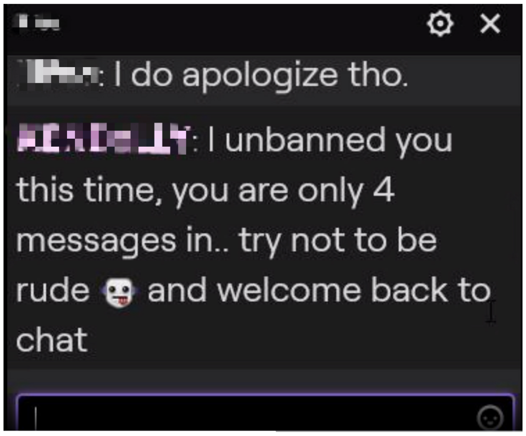
date and subscription status in the channel, by quickly clicking on the username. Subscription meant that users paid a monthly fee to support the streamer in the channel, indicating the enjoyment of the content and the contribution to the streamer. P8 (F, 18) noted, *"People typically don't throw money at the people they want to mess with and make a bad day."* Thus, the subscription was a good reflection of a user's intent. Through the observation, we asked P18 (F, 20) why she timed out a user. According to Figure 6.2, the user typed "fresk is bad" and got a 10-minute timeout. Fresk was the streamer's name, and the mod considered it a personal attack: *"What I do to judge is I check if the person is a sub. As you can see, he's not a subscriber as well, so I know he's not really joking."* In P18's explanation, subscription status could also be reflected by the subscriber badge alongside the username. In this example, the username had no badge and got a timeout.

Mods also checked the following date to distinguish the regular from new users. Following a channel was free and an indication of a user's interest in the stream. After following the channel, users could send messages in some follower-only chat channels. P2 (M, 19) explained, *"I would say I would click on their names, and I'll check their following age and see if they're following the person that they hosted from or see that it's a random person who just saw because of the high number of viewer count, and if it was a toxic message, and they were just following the person, then I would time them out."* In P2's sense, a short following time with toxic content suggested the user's intent to harass others. Overall, channel status indicated the loyalty and interest of the community. Mods had the mental model that users with long following time and subscription periods were valuable community members and less likely to be sanctioned.

Account Status Account status refers to the user information and activities on the platform (community). Four mods stated that they would check account age that determined the user's length on this platform by clicking on the username. They consistently agreed that the account age was a good indicator of a troll account or throwaway account. We observed P3 (M, 23) checked a user's account information after a user typed "wtf" in the body painting channel and asked him why he checked and what he was looking for: *"Usually when people say something like that, I immediately think, okay, how old is the account? Do I need to ban them? But I don't believe he said anything else, so I kind of just let it go."* According to P3 and our observation, the account was created in 2019, so it was an old account, and he also checked the message history, indicating this was the first message of the user. Thus, he *"let it go"* and gave the violator another chance. He further explained why he considered account age very important for his moderation: *"If somebody makes a throwaway account, they can just make a new one right after. They don't have to worry about, then bans technically won't even matter, but if it's an older account, more often than not, I'm less likely to ban them."* Typically, throwaway accounts were created in a short time; thus the account was new and usually untrustworthy. P7 (M, 18) also reported similar logic: *"I can check the account creation date, so I'll know if this person just made their account two hours ago. Chances are it's just like a troll account. So there's no harm if we just ban it, but that isn't to say if the person's had an account for six years, they wouldn't do something like that. So it's called context-based. I would say I'm harsher toward accounts that were recently made because I feel like you're making an account to troll, like you're going to get banned."* Similar to channel status, account status with longer age was considered more valuable to the community. Differently, channel status only reflected the activities in a specific channel, but account status reflected the account activities on the platform. The platform contained thousands of different channels. A poor channel status did not necessarily indicate a poor account status and vice versa. Mods relied on both.

Sequential Evidence Sequential evidence refers to information that indicates the sequence of the act (e.g., chat messages with timestamps). Mods reported scrutinizing a user's message history to 1) gain context of a specific situation or a user, 2) identify the behavioral pattern, and 3) review moderation history with timestamps. The difference between action evidence and sequential evidence was that action evidence emphasized the single action reflected by the chat message, while sequential evidence emphasized the actions in the sequence.

Chat Context (Recent Chat) Mods often collected chat history to gain the context of a specific situation. Checking chat history facilitated their moderation actions. P19 (M, 26) stated, *"We can see their past messages. So sometimes we'll look at that and see what started the argument and like, why were they arguing with each other, why were they talking to each other?"* Similarly, P15 (M, 31) expressed that he usually *"scroll up in the chat and find out what the context is."* By comparing chat history, mods resolved the issue fairly. P13 (M, 29) described that he often went back to the message history to *"compare"* everyone's message to gain *" a little more context"* and resolved the issue. Mods also used chat history to gain an understanding of the users. P19 (M, 26) explained how he used previous messages to know users' temperament: *"I look for what they say because I never really like just anything that jumps out as toxicity or overall negativity. That's not worthy of me banning them just because of what they've previously said, but it does let me know what kind of*


[redacted]

[redacted]: I do apologize tho.

[redacted]: I unbanned you this time, you are only 4 messages in.. try not to be rude 🤪 and welcome back to chat

Figure 6.3 A screenshot of Whisper conversation between a viewer and a mod on Twitch. After a violator apologized, the mod unbanned them with an explanation and maintained the "violator" in the community.

temperament they have." According to P19, checking previous messages was a way to understand the *"temperament"* of potential violators.

In some cases, though the users seemed to perform well, they might break the rules later. Several mods reported that if they saw single-letter expressions, which were indicators of potential spam and personal attack, they started to check the recent chat history. P15 (M, 31) said that violators used one-word messages to spell out something inappropriate, and they watched out for this type of message regularly. In rare cases, mods would like to sanction first if they lack context. P14 (F, 28) explained her moderation preference (sanctioning first, then revoking if the violator explained) and shared with us a video clip of a Whisper conversation with the violator (see Figure 6.3). She revoked the sanction after the violator sent a private message to explain the situation and apologized. She kindly reminded the violator to be careful and explained that the violator had only four messages in the chat history.

Behavioral Pattern via Chat History Mods used chat history to identify the behavioral pattern of violators. P17 (M, 21) described his identification of a recurring troll: *"If they keep saying the same shit over and over again, like someone asks a question, right? I answered, and they ask again. That happens too many times, and* *then I click on his name, and I checked: ‘Wait, is he like spamming the question or not?”’* In Figure 6.1b, P17 (M, 21) checked a user chat history after the user typed “crybaby” and explained, *“I see something in chat, and I’m like, okay, is this guy toxic like usual? Is he usually toxic? Is it like a one-time occurrence, right? So I’m looking [at] his chat log to see.”* P17 found that there were only five previous messages and that this was a one-time occurrence, so he decided to let it go. The behavioral pattern not only showed what has happened but also predicted what could happen (P13, M, 29).

Moderation History Moderation history included the messages being banned and timed out. Some messages were under the same moderation action, and some were even the same. These messages were repeated offenses instead of generally repeated behaviors. P12 (M, 21) described how he checked the repeated offensive messages in moderation history as references: *“I see if they’ve been banned before because if they’re a repeat offender, I don’t even think about it. They’re just going to get banned again. Things like, have they sent where they banned for the exact same message before? Where they timed out for the exact same message before in a different stream? That sort of thing.”* According to P12, mods checked *“the exact same”* offensive behaviors through the moderation history. The same messages guided them to sanction the violators. Interestingly, mods indicated that they also referred to evidence from different streaming channels.

Generally, among the three types of evidence, action evidence works as a trigger, ownership evidence as a start, and sequential evidence as a supplement. After seeing a negative message (action evidence), instead of commonly filtering and blocking as moderation strategies, mods first use visual cues such as username and badges to form an impression quickly (ownership evidence), then checking account information to make sure whether this is a first-time violator (ownership and sequential evidence).

If they lack the context, most of them would like to give users another chance and track with close attention (sequential evidence), waiting for more evidence to understand the context and intent.

6.4.2 Evidence Collection

To answer the second research question: how mods collect these types of evidence, we found that mods collected these types of evidence in five different ways, including documenting, co-experiencing with viewers as inference, collaborating with moderation teams (across channels), gaining knowledge from users by staying in the community for a long time, and relying on moderation tools.

Documenting A few mods stated that they shared a spreadsheet containing violators' information with notes in the moderation team. They also did cross-channel documenting, which meant several channels individually documented the violators and shared with others. Documenting was a way to mainly collect ownership evidence. P15 (M, 31) described how the information collected on a Google document helped him gain context of the violation: *"We have a shared Google document. It has a list of not finding the word, but people that have caused problems in the past for timeout or ban or whatever. If I lack context in a situation in that community, then I can go to that spreadsheet. I can search for that person's name, and I can see if they've been a problem in the past or if this is their first infraction."* Furthermore, P15 stated that they also shared the document across different moderation teams so that other channels could pay close attention to these violators: *"We have a sheet that is just for known troublemakers, so moderation teams from other streams will see. These are the people that we had issues with. Here are their usernames so that you can be aware, and then somebody comes in, and they start saying something that they might seem to be innocent at first, but we know based on the information from another moderation team that this is someone who has been a problem in the past, so we can watch out* *for them if they start to go down a path of being a troll or whatever."* According to P15, mods applied external platforms that were not initially designed for moderation to collaboratively moderate, either within the channel or across different channels.

Co-experiencing with Viewers as Inference Some mods reported that they were viewers and watched other channels that streamed similar content to their moderated channel. Similar streaming content attracted similar types of viewers. Thus, they knew the background and actions of violators in other channels. When these violators came into their channels, they recognized them. For example, P7 (M, 18) stated, *"There're also people from other streams that come in, I know from their stream, their respective place."* P13 (M, 29) described how he recognized viewers from other streams through ownership evidence: *"Some people have some pretty weird names, right? You can kind of see. When you see the kind of stuff that you don't really think is right, you kind of subconsciously remember it a little more."*

Collaborating with Moderation Teams (Across Channels) Many mods also reported that they collaborated with other mods in either the team of the channel or teams across channels in mainly three ways (asking other mods' opinions within the channel, cross-channel log check, and multiple channel moderation). The nuanced difference between cross-channel log check and cross-channel documenting was that log check included all chat history while documenting only included violation behaviors. The difference between co-experiencing with viewers as inference and multiple channel moderation was that mods were viewers in other channels in the former situation and were moderators in other channels in the latter situation.

Some mods asked other mods' opinions, like sending a *"screenshot of the message or log"* (P18, F, 20) to others when they lacked background information of a particular viewer, in line with prior work that mods had group discussion during the moderation process [133]. A few mods stated that they had a collaboration with other streamers and could conduct cross-channel log check through third-party platforms. P14 (F, 28) described how her team applied a third-party platform called Overrustlelog [5], a public chat log website for Twitch channels, to collaborate with other streams: *"I know all of our mods, we do this like, for example, a lot of people don't like XXX. She's another streamer, and we've had to ban a few of our people that went over to her chat to be toxic. We know this because we saw in her Overrustlelog, so it wasn't just hearsay ... We've had to cross ban people that got from our community had gone over to her chat to be douchebags, and so we'll ban them in our chat."* According to P14, mods sometimes moderate not only violators within their channels but also users who were considered violators in other channels though they did not break the rule in their channels.

A few mods also noted that they moderated across different channels sharing similar viewers, and the viewers' behaviors in other channels could be indicators of their decisions of the current channel. P10 (M, 18) shared his experience moderating two streams with the same content: *"When both streamers are live, or when they were streaming together, like playing this together, you would have viewers in one chat that are toxic in one chat that would obviously be toxic in the other. Since I'm a moderator for both, it's kinda clear. I remember viewers from one chat that break the rules a lot."* Some violators kept the same username across different channels; mods easily remembered their names. P8 (F, 18) stated: *"A lot of the time people that are there to cause problems, they don't change accounts. They just keep the same name, so what you find in maybe one person's stream, you might ban them, and then you might see their name a few hours later, and you'll go, I remember that name."* Both P10 and P8 in common described they remembered violators' names in a short time, either at the same time or *"a few hours later."* Team collaboration mainly relied on usernames as references and violation history to gain context. Mods also expressed the challenge

[5] `https://overrustlelogs.net/`. Retrieved in March, 2021

of identifying violators if they completely changed their usernames across different channels.

Gaining Knowledge from Users by Staying in the Community for a Long Time Some mods stated that they had been in similar communities for a long time and recognized viewers through frequent seeing. P15 (M, 31) said, *"I've been in this channel for seven years at this point. You spent time in channels over time, you learn the regulars, you get to know them, and you recognize them."* Similarly, P6 (F, 34) said some users actively appeared in Twitch chat and Discord channel to interact with others, and mods *"kind of know how long they've been around."* Combining the frequent seeing of usernames with other evidence helped mods figure out the intent of users. P16 (M, 24) explained, *"So they've subscribed to XXX, I think that's a five or six-year badge, so that's a lot of money to give to XXX. I've seen their names a lot. I can tell that they like XXX, So if that person types the same message, I fucking hate you, and I'm going to probably understand it as 'oh he's jokingly hating the person."'* According to P16, mods combined account status, badges, and frequently seeing users to interpret users' behaviors, finding out that this user was *"jokingly hating the person."*

Relying on Moderation Tools Most mods applied various bots in addition to the AutoMod offered by Twitch to facilitate the moderation process. Many mods applied third-party tools, such as Better Twitch TV and FrankerFaceZ, to customize moderation action, similar to prior work [15]. As shown in Figure 6.2, there was a list of customized buttons in front of the username. At the same time, tools allowed mods to collect various types of evidence such as account age, channel status, and message history in the channel. For example, P5(F, 27) sometimes *"go through and check their profile"* to determine throwaway accounts with the assistance of moderation tools. They mainly collected ownership evidence and sequential evidence, Figure 6.1b showed a typical interface of Twitch AutoMod. This account was created in 2018, indicating that it was an old account. It followed this channel in 2020, several months ago. Bans and timeouts were "0," indicating it might be a good user. Moderation tools provided the necessary information to help mods form the first impression on users quickly.

6.4.3 Types of Violators

To answer the third research question, 'What are the types of violators that mods perceive?' we identified five types of violators. Moreover, *"racist"* and *"sexist"* were commonly mentioned by mods with a consistent attitude toward sanctions. They are easily recognized via action evidence and sequential evidence, with the assistance of ownership evidence. Mods would ban them without further consideration. We present the other five types of violators reflecting mods' complex attitude and decision-making process.

Violators Performing Malicious Mischief Criminal mischief, also called malicious mischief, refers to behaviors intentionally damaging another person's property in criminal justice. Several mods reported a type of violator who randomly came into a channel to cause trouble and intended to disrupt the community. For example, P15 (M, 31) said, *"You have people who come in ,and they just want to be malicious. They come in specifically to be disruptive. They come in specifically to cause an issue, to force the mod team to do something."* Similarly, P6 (F, 34) expressed that this type of violator wanted to see the anger from the streamer: *"I think they just want to get a rise out of the streamer. They want the streamer to kind of fightback there."* This type of violator took advantage of the anonymity and pseudonymity of the Internet and obtained excitement from the mischief. P13 (M, 29) said, *"Maybe they just appreciate the anonymity or that, and they're just like, 'hey, we can haha, we can get a rise out of people if we do this."'*

Attention and Reaction Seekers Many mods mentioned that a type of violator was the attention and reaction seeker. Unlike violators performing malicious mischief, these attention and reaction seekers did not initially try to cause trouble. They competed for recognition mainly through sarcasm and troll and for popularity through self-promotion.

Attention Seekers Needing Recognition Some violators wanted to *"get recognition from a streamer"* (P6, F, 34). *"They're trying to get people to notice them, to validate them and their actions, so it's not always because they disliked the stream or they disliked the viewers. It's because they need to be seen and recognized, and they have the need to be validated,"* said P15 (M, 31). These violators broke the rules because they had the desire to be recognized by others. Once their needs were recognized and fully fulfilled, violators might *"turn to normal people"* (P19, M, 26). However, the overwhelming messages made the streamer not recognize them. P4 (M, 18) explained that *"everybody wants attention"* and said, *"Because they like watching the streamer, so they want attention from the streamer, reading the question, answering it or saying hello to them. It's a personal connection through the screen."* In P4's sense, attention for recognition was considered a strong personal connection with the streamer. Massive viewers wanted to be recognized by the streamer, thus forming completion. In order to stand out, some violators attempted to be sarcastic or make trolls. P3 (M, 23) suggested that some sarcastic jokes were in the *"grey area."* Thus, mods needed to put it into the context of the conversation to interpret its meaning. For example, P19 (M, 26) told us that they could usually differentiate whether it was a sarcastic or toxic comment: *"We can usually tell because there'd be other types of comments in there. There'll be conversational comments with other chatters. There'll be other statements about stream... Those would normally be considered a toxic comment, then put into context of what they love, what other things they said,* *and you realize it's potentially not toxic. It's potentially just sarcasm."* According to P19, mods mentally categorized comments into different types and applied the chat history as a context to interpret the underlying meaning of the messages. Mods relied mostly on sequential evidence to make the judgment. In extreme cases, some violators experienced mental health issues in offline life and started the *"psychological cry for help"* (P15, M, 31) online. P16 (M, 24) explained that he believed the violators were not negative offline, and most violators did not have an outlet to let all anger and depression out in real life, so they came to online communities.

Attention Seekers Seeking Popularity Another type of attention seeker was violators who wanted to gain popularity by promoting themselves in other streams. P4 (M, 18) described that some streamers (competitors) in small channels went to the big channels to post advertisements and *"make as much noise as they can"*. Gaining popularity was the main reason, and *"attention, popularity, intention kind of go hand in hand."* P4 added, *"They can make a disturbance and say, you know, go follow me on this, on their social media sites, or they'll shout themselves out in front of thousands of people in chat. That's obviously not acceptable."* Similarly, P6 (F, 34) noted, *"You have the kind of attention seekers who will hop in and be like, 'hey, look at me. I'm a streamer to those.' We don't like those. I don't want other people advertising, so we get rid of them."* P4 and P6 indicated this type of attention seeker were not *"acceptable"* and would like to *"get rid of them."*

Immature Juvenile Four mods mentioned juvenile as a type of violator and usually treated it differently. P8 (F, 18) explained that she moderated in years and could *"pick up on the pattern"* to identify juvenile violators through messages and tones they used: *"He tends to talk in caps with very bad grammar and you can kind of look at that and go, 'Oh, that is more than likely a little kid rather than a problem.' He also thinks like the most random things are funny ... You can tell that they think it's* *really funny, and that tends to be more of childish humor. It doesn't mean everyone with that humor is a child, but it leans more toward being a child."* According to P8, juveniles preferred using capital letters with bad grammar. The language pattern also indicated that juveniles and adults had different senses of humor. P15 (M, 31) supported P8's explanation: *"They'll use the letter U instead of the word YOU, they'll use the letters UR instead of YOUR, and they'll do a lot of things like that to make it abbreviated, to use what people called 'tech speak."'*

After mods identified the pattern of juveniles, they preferred communication to sanctioning. P8 said, *"We try to talk to them more than actually take action against them because we're trying to help them understand why what they're doing isn't proper."* P15 further shared an example: *"He went straight to saying very inappropriate things about the streamer and [the streamer] talked to him, asked him what was going on, and I ended up stepping in and talking to him, asked him if he needed to talk, ended up talking to the kid for a couple of hours that night and come to find out his parents were going through a divorce, and his dad had abused his mother that day and then left the house, so he was upset. He didn't know how to properly vent his feelings, and his way was to go onto Twitch and try and be a troll. So ended up talking to him for a few hours, and then he became an active member of the community for a couple of years after that."* According to P15, consistently, mods reported that some violators experienced mental health issues and used online communities to vent emotions they suffered from offline life. Mods tended to have a strong tolerance for juveniles' violations and would like to help. In this sample, communication helped the juvenile and transferred the violator to an active community member.

Repeated Offenders with Contributive Participation Mods stated that some violators were toxic regulars but also active community members. These violators kept breaking the rules, accepting punishments, still staying in and contributing to the community, and breaking the rules later. These violators were *"stubborn"* and *"unable to adapt or change"* (P12, M, 21) but valuable community members. P16 (M, 24) said, *"The part that makes me like them is that they do actually interact with the chat room. It's like they talk to each other. They talk to the streamer. Occasionally with frequency, they will break the rules. It's like a very nice criminal that you consistently arrest, but they're always respectful to you. They're respectful to the content of the streamer. They're respectful to the streamer. They're respectful to everybody else in the chat room, but they just have this habit of getting in trouble."* According to P16, some violators having the *"habit of getting in trouble"* are active community members and *"respectful"* to the community. These violators are considered *"nice criminals"* because they accepted the mistakes they made and the sanctions that they were given with no intention to leave the community. Some mods had mixed feelings and concerns about the punishment for this type of violator. Generally, they sanctioned them differently, considering their contributions. P9 (M, 19) shared his experience moderating "active" but also "toxic" viewers: *"So it's very difficult to decide how we're going to deal with them because they're still a very active part of the community. They're contributing a lot to the community. It's just like, occasionally they make mistakes that are against the rules, but we punish them differently because they're adding a lot to the community and they're like helping. So it kind of gets hard to figure out what sort of punishment we're going to give them."* According to P9, mods sanctioned the repeated offender and repeated offenders with active participation differently and experienced difficulty in deciding the sanction level to this type of violator.

Aggressive and Hostile Attackers Another type of violator was viewers who were aggressive and hostile. This type of violator could be easily triggered to start harassing or attacking others. P12 (M, 21) said, *"We have viewers who come in,* *and they do stuff they're not supposed to do, and then they get timed out for it, not necessarily banned, but then they get aggressive. So they'll either in my whispers, 'why'd you time me out? You're a piece of shit. Like kill yourself'... or after the 10 minutes they'll come back and chat, 'Wow, your mods are absolutely trash, blah, blah, blah, like fuck you."'* In this case, the violator was not satisfied with the moderation and started the aggressive behaviors through either Whisper or the public chat to attack the mod. Some violators who got banned in Twitch communities targeted other relevant communities to continue the attack. P14 (F, 28) shared an example of a violator posting on the subreddit of the Twitch channel to accuse that mods abused the power of banning people, and then the Twitch mods and the violator started the argument on Reddit. In other cases, if mods understood violators' personalities and knew their intent, they might allow it. P6 (F, 34) said, *"I know we've had one person that's been a regular, and she'll often do the backhanded threat of 'I will cut you'. We know she's not going to, but it's more of that feisty spirit more than anything. So it's like, yeah, we know she's not really going to attack this person."* According to P6, though this violator threatened other viewers, the moderator knew this violator and considered the violation behavior not serious enough to warrant punishment, whereas someone else who said the same thing might have been subject to a different type of sanction.

6.5 Discussion

We use criminal profiling as a lens to guide us to understand the mental model of mods who have a non-expert profiling background when they deal with potential violators. Mods work as both evidence collectors and profilers in the moderation process. We find that mods mainly collect three types of evidence in five different ways. The five methods of collecting evidence mainly rely on individual experience and collaborative work with limited technical support from the platform, mostly collecting ownership evidence and sequential evidence. After the evidence collection, mods unconsciously fit violators into mainly five types and apply different moderation strategies.

We clarify that the mental model in this work consists of two parts: the first is about collecting and using different evidence; the second is about the types of violators requiring different moderation strategies. The pattern of evidence types and collection might generalize to other online communities that aim to thrive via extensive effort in the moderation process. The different affordances of platforms might cause the process to be a little different. For example, on asynchronous platforms such as Twitter and Reddit, users' activities such as posts and replies are saved under a user's profile, making the evidence collection process comparably easy. Content removal and banning users are easy and sometimes can effectively decrease toxicity from existent users but force other users to migrate to other platforms [25]. The types of violators identified in this work show the complexity of users' behaviors. These types provide community administrators an alternative to consider punishment if they aim to maintain community members. Meanwhile, commercial moderation teams who work for social media and news sites might integrate the mental model into the moderation process and use it to restrain severe sanctions for first-time offenders.

6.5.1 Platform Design and Affordance Make Profiling Go Beyond the User's Profile

A user's profile often contains registration information and account activities. Prior work has explored how users on social media sites curate self-presentation to maintain social relationships with other users through different profile elements, such as a profile image [162, 163], the about me and interest [69, 99], and the location field [72, 153]. Account activities under a user's profile provide cues to understand the user. For example, peers in the open-source community form impressions about other users' expertise based on the history of activity across projects and the successful collaboration with key high-status projects [108].

In live streaming communities, we find that many mods frequently mention they check the account status and channel status because of the limited information on a user's profile. The interface of a user's homepage is initially designed for those who will be streamers. We speculate that the design discourages viewers from filling in the relevant information. In addition, different from posts and feeds on social media sites like Facebook and Twitter, activities such as message histories and replies are not stored under the user's profile from the user's end because of the synchronicity and ephemerality of the "live" affordance. In other words, after the streamer closes the stream or the user leaves the channel, the user cannot store or see the message history in the channel anymore. Once the users leave and come back, the message history is erased and displayed from the time point the user gets in. In addition, mods have to apply tools to log chat histories of a user only in the specific micro-community/channel. Mods in the current micro-community cannot see the message history and violation in other ones. Mods also do not have access to log data of other micro-communities. The limited information on a user's profile and the challenges of acquiring other information compel mods to seek other methods to collect evidence beyond a user's profile and across various micro-communities. Thus, understanding evidence collection is essential to figure out the profiling.

6.5.2 Profiling as Part of the Moderation for Community Growth

Different from the goal of criminal profiling for crime capture [47], in online communities that include thousands of micro-communities, the goal of violator profiling is to avoid punishing users, even help users in some cases, to grow the micro-communities. In our observation, mods rarely directly ban users only based on the content. Even when they do so, they can easily revoke after users express remorse and apology. On Twitch, mods frequently play roles to facilitate the community, such as facilitator, mediator, and adult in the room [129]. Overall, mods are willing to go the extra mile to retain community members.

Fairness and Justice In criminal justice systems, retributive justice suggests sanctioning violators with proportional punishment for their violations [21] and is predominantly applied on commercial platforms. Most volunteer mods in user-governed micro-communities show a preference for restorative justice, involving the repair of justice by bringing stakeholders such as violators, victims, and mediators together to acknowledge and remediate harm [155]. Prior work shows that retributive justice is not the most effective measure to promote reconciliation, and restorative justice can potentially complement it to initiate and boost reconciliation [32, 10]. Mods in live streaming communities often work as facilitators to mediate the conflict in the chat, such as asking users to change or stop a particular behavior and helping users with trouble in offline life. Profiling as part of the moderation process in live streaming communities shows an example of the application of restorative justice to users, supplementing recent work appealing a restorative justice to support targets of harassment online [127].

The complex behaviors for each type of violator indicate the same standards of punishment to these violators are considered unfair and unjust. The one-fit-all approach will fail and drift away from these potentially valuable users [127]. After profiling, mods identify the types of violators who need help or unintentionally break the rules. Mods choose to communicate with and take care of them instead of outright punishing them. The caretaking and restorative approach make these one-time or one-day violators become loyal community members later. Accordingly, sanction after profiling could potentially increase the perceived fairness and justice in these spaces.

Bad Act and Bad Actors Profiling allows mods to ascertain the user as a bad actor not only based on the bad act at the scene. Our work supplements prior work arguing that moderation should consider the context [22] and reveals some more sophisticated scenarios. Mods consider not only the content and context but also the violator's intent and characteristics, sometimes their experience, into moderation. Many mods describe they apply the other channel violation as a reference of moderation action in the current channel; in rare cases, mods rely on what happened in other channels as a way to understand a user's personality, not a reference of sanction.

Automated moderation systems heavily rely on the content and consider the bad act as a violation and sanction bad actors. Our results show that though mods recognize the "bad" actors, it is difficult for them to assign the punishment in some situations. For example, mods express that they weigh the violators' contribution and tend to have more tolerance to repeated offenders with contributive participation. Notably, many mods consistently express emotional and social support to immature juveniles and would like to talk with and educate them. They also sanction similar behaviors differently for other types of violators. For example, knowing the aggressive and hostile attackers' personalities and intent is critical for mods to decide approval or ban; attention seekers seeking popularity are directly banned, but those needing recognition depend. Profiling in the moderation process attempts to decrease the bias and discrimination created by the automated moderation system [62, 11] and allows mods to distinguish the bad actors from the "bad" act.

6.5.3 Implications and Recommendations

We propose designs to facilitate collaborative and individual violator profiling and to integrate violator profiling into the moderation system that combined automated and human processes.

Build a Mechanism for Collaborative Profiling Mods are collaboratively getting rid of violators, collecting violators' information across different micro-communities via external tools or platforms not essentially designed for profiling. According to the official website of OverRustleLogs, it was shut down in May 2020 at the request of the Twitch Legal team because of privacy concerns. However, the internal tools only work for a specific channel. We suggest the platform develop a mechanism that allows all mods at the micro-community level to list violators and to share the information with other mods at the community level. The pseudonymity of Twitch helps reveal more information of violation while keeping violators' real identities safe.

Cross-channel collaborative moderation also indicates the possibility of cross-platform moderation, Tech giants (Facebook, Microsoft, Twitter, and YouTube) together established the *"Global Internet Forum to Counter Terrorism"* in 2017 to coordinate content removal about "violent terrorist imagery and propaganda" [66]. However, there is little or no collaboration about dealing with daily online harassment. We propose a mediated system that allows different online communities to document and share violators' information to the commercial moderation teams or volunteer moderation teams. Though commercial moderation teams work behind the scene and are managed by the platform [124], which can protect the violators' information privacy while allowing mods to deal with the violation, we don't know how to keep the boundary between privacy and profiling in the volunteer moderation teams, requiring further investigation.

Facilitate Individual Mod to Profile Violators Some recommendations to facilitate individual profiling should be highlighted to supplement collaborative profiling. First, we suggest a mechanism allowing mods to label and tag violators manually. Recent work has developed prototypes to use algorithms to analyze the message history to automatically label users [78] and summarize messages as key points [161]. Our findings reveal the complexity of violators' characteristics. Thus, we suggest integrating a mechanism (either developed by Twitch or third parties) containing a database with pre-defined personality traits and violator types in criminology and psychology. These labels might help mods scrutinize factors that are not achievable by algorithms and allow them manually tag violators.

Second, we suggest a feature allowing mods to trace username change. Mods explain that sometimes they can remember the usernames or recognize the users, but the vague memory and change of usernames increase recognition difficulty. The process is primarily supported by social, not computational practice, making the recognition very random. The current AutoMod allows mods to log the message history of users. The pseudonymity of online communities encourages self-disclosure and free speech [150] but also increases the profiling challenge. We suggest developing a feature in the moderation tool that can trace the username change history across different micro-communities. These designs align with the current Twitch moderation mechanism, which is only visible to the streamer and mods to facilitate the moderation process.

Resource for Training and Educating Mods Prior work shows professional profilers can produce a more accurate prediction of an unknown offender, comparing to other groups [96]. We find that mods, as non-experts in profiling, own much power to sanction violators, and the process sometimes is pretty subjective, varying from person to person. Platforms might offer resources for training and educating mods to avoid false profiling, such as making online video tutorials to explain the importance of profiling and integrating the components into the moderation guideline to show mods how to profile step-by-step.

6.5.4 Limitation and Future Work

This work suffers several limitations. First, the data collection is from a single platform — Twitch, which is different from other asynchronous communities. Future work should do cross-platform research to validate the findings. Second, our participants were mainly in Europe and North America, but live streaming service is also booming in Asia [106]. Future work can apply our findings in a cross-cultural context. Third, our participants are mods who are willing to share the video and content; thus, we may have recruited mods who are more inclined toward restorative justice. We do not know the justice preference of the mods who are unwilling to share content. Moreover, whether the recording task affects mods behaviors needs further investigation. Fourth, though we show profiling violators as a phenomenon in live streaming communities, we can not answer questions like how frequently mods use profiling in the real-time context. Future research can apply quantitative methods with log data to explore this question. Additionally, streamers' characteristics (e.g., gender, age, preference) and the channel characteristics (e.g., content categories, community size, clarity of rules) might also have significant effects on how and when mods will choose to use profiling during the moderation process. Future work can incorporate these characteristics into algorithmic models to investigate their relationships. Last, we don't know if the profiles that moderators create are accurate representations of the violators, as we only focused on moderators' thought and behavioral processes in constructing these profiles. Future research may want to see if these profiles are accurate assessments.

6.6 Conclusion

In this work, we aimed to understand how volunteer mods on Twitch create profiles of violators before they decide on what action they will take with the violator. We found that profiling improved mods' understanding of violators, and they engaged in complex practices of evidence collection and documentation to create these profiles. These practices happened not just within one community but across different Twitch communities as well as on different platforms.

Generally, instead of sanctioning violators, mods preferred to go the extra mile to integrate the violators into the communities. Though they had to sanction some violators, the profiling led to different sanction decisions. We also found that mods across different micro-communities collaboratively worked on violator profiling because of the limited information in the user's profile and limited technical support from the platform.

CHAPTER 7

CONFLICT WITH MANAGEMENT IN THE MODERATION TEAM IN LIVE STREAMING COMMUNITIES

7.1 Introduction

Norms and rules play critical roles in regulating human behaviors in many online communities [22]. Different platforms might apply different moderation philosophies to enforce rules and norms. For example, social media giants like Facebook and Twitter may apply commercial moderation, hiring contractors to moderate with formal guidelines and instructions [59, 124]; other platforms like Discord and Reddit, containing many different micro communities apply community moderation, relying upon the micro communities to select their community members as volunteer moderators (mods) to govern their users [133].

For platforms applying community moderation, communities often develop their own rules through discussions among volunteer mods and community members [100, 110]. As guardians to manage and grow the communities, volunteer mods have received much research attention (e.g., [59, 124]). However, they also experience conflict in the rule development process regarding what is acceptable and how to punish violators. High levels of conflicts or specific types of conflicts can threaten the speed of decision-making, hinder implementation [82], and even threaten the continuity of communities [118]. Though leaving one community and creating another one is the straightforward way to handle conflict in the moderation team (e.g., Reddit mods leaving a subreddit to create a new subreddit [44]), are there other ways to handle conflict in the moderation team? Moreover, are there any other types of conflict? How are the conflict and management styles associated with mods' commitment to the community? Despite that conflict management has been well-established in face-to-face communication (e.g., [41, 122, 147]), and to some extent, in online collaboration systems such as Wikipedia [93, 94], GitHub [77], and other open-source software development communities [50, 151], it has been under-explored in community moderation teams.

In live streaming communities, the moderation team is formed and led by the streamer, consisting of both the streamer and volunteer mods. Mods are usually motivated by helping the streamer or the community in general to have a good experience [157]. Streamers can easily appoint other users as mods with permission or revoke mods' status. Even though prior work in HCI and CSCW has documented the application of live streaming in diverse domains from the streamer-viewer relationship perspective (e.g., [68, 105, 20]), the streamer-moderator relationship has received relatively less scholarly attention. On live streaming platforms, the micro community (called "channel" on Twitch) is streamer-centric; different streamers employ different rules to meet their expectation. However, many channels don't have clear rules or even have no rule at all [14]. Lack of clear guidelines often leads them to disagree about what is acceptable and what decision they should make.

This research focuses on community moderation on live streaming platforms and explores the triangle relationships among conflict types in the moderation team, mods' conflict management styles, and mods' commitments to the streamer. We contribute to understanding mods' conflict management during the moderation process in user-governed online communities and providing insights to micro community leaders and mods who seek to handle conflicts effectively to grow the micro community.

7.2 Related Work

We first review and summarize the types of conflict. Then, we discuss online community commitment and develop hypotheses among conflicts and commitments. Next, we introduce conflict management styles and develop our research questions exploring their relationships with different conflicts and commitments.

7.2.1 Organizational Conflict and Conflict in Online Communities

Conflict is *"an interactive process manifested in incompatibility, disagreement, or dissonance within or between social entities"* [122]. In this study, we focus on the intragroup conflicts within the micro community of live streaming — the moderation team of each channel. Early organizational research generally divides conflicts into two types: task conflict (disagreement relating to task issues) and relationship conflict (incompatibility relating to emotional or interpersonal issues) [65, 80, 2]. Later evidence has suggested another type of conflict — process conflict. Process conflict happens when group members disagree about the logistics of the task, such as the delegation of tasks and responsibilities [82, 81]. Normative conflict is defined as a perceived discrepancy between the current norms of a group and an alternative standard for behavior and often arises from inconsistencies between aspects of identity [120]. Normal conflict is associated with organizational rules and identification [36] and online community rules and norms, such as policies, governance structures, and ideology [50].

Some research has explored the source of online conflicts and different types of conflicts in task-oriented online communities, highlighting the importance of understanding conflicts and their impact on community development [76, 50]. For example, in open source communities, task conflicts happened between professional and voluntary programmers in that they had different viewpoints and backgrounds of the projects and programming; affective conflicts happened in that people worked globally with different cultures and languages [151]. Other work has examined the common pattern of conflict from the ground and extended conflict types such as procedural conflict (how to do the task) and normative conflict (what norms to follow to do the task) and has explored how task, procedural, relational, normative conflicts intertwined [50]. Inappropriate handling of these conflicts can cause poor group outcomes such as poor performance, dissatisfaction, and member attrition [50, 112].

These specific conflicts still fit into the four types. In this study, we apply the four types of conflict to live streaming communities.

7.2.2 Online Community Commitment

While plenty of work has explored how different conflicts affect various group outcomes such as productivity, effectiveness, satisfaction, and propensity to leave (see meta-analysis by [40]), little work has systematically explored how these conflicts affect community commitment as a type of group outcome (the feelings of attachment to the goals and values of the community [34]).

There are three types of commitment in organization research, including affective commitment (emotional attachment to the organization), continuance commitment (awareness of the costs to leave the organization), and normative commitment (feelings of obligation to remain with the organization) [1]. Though commonly applied in the organizational context, commitment research originally explores why volunteers' dedication varies at nonprofit organizations [7], making it a particularly appropriate theory base for understanding an individual's voluntary behavior in online communities [6]. In live streaming communities, mods are either motivated by building a personal relationship with the streamer or by helping the streamer to grow the community [157]. Thus, we adopt the three community commitments into volunteer mods' commitments to the streamer: affective commitment to the streamer, continuance commitment to the streamer, and normative commitment to the streamer.

Prior work suggests that relationship conflict in the group is positively associated with an employee's propensity to leave a job and satisfaction [112], and affective commitment is positively associated with extrinsic and intrinsic satisfaction [107]. The literature has consistently suggested that as the relationship conflict increases, the affective commitment to the group decreases [135, 80]. Similarly, we assume that the perceived relationship conflict in the moderation team is negatively associated with mods' affective commitment to the moderation team (streamer and other mods). However, we don't consider the commitment to other mods in this study. Accordingly, the following hypotheses exploring the relationship between conflicts and commitments are developed:

- **H1a:** Perceived relationship conflict in the moderation team is negatively associated with the mod's affective commitment to the streamer.
- **H1b:** Perceived relationship conflict in the moderation team is negatively associated with the mod's continuance commitment to the streamer.

Task conflict can decrease group loyalty, workgroup commitment, intent to stay in the present organization, and job satisfaction [83], and is detrimental to group functioning when members conduct routine tasks [80]. Similarly, we assume that the perceived task and process conflict in the moderation team is negatively associated with the mod's commitment to the streamer.

- **H2a:** Perceived task and process conflict in the moderation team is negatively associated with the mod's affective commitment to the streamer.
- **H2b:** Perceived task and process conflict in the moderation team is negatively associated with the mod's continuance commitment to the streamer.

In the organizational context, normative conflict with organizational rules decreases employees' affective and normative commitment to the organization [36]. Prior work suggests that strongly identified members are likely to challenge community norms when they experience conflict between norms and important alternate standards for behavior, in particular when they perceive norms as being harmful to the community [120]. As a user-governed online community, the streamer and mods develop the rules with other community members' feedback. If mods perceive an inconsistency between what is expected and the community's rules, they might show a sense of obligation to the streamer and provide suggestions to the moderation team.

- **H3:** Perceived normative conflict in the moderation team is positively associated with the mod's normative commitment to the streamer.

7.2.3 Conflict Management in Online Communities

Conflicts can be both constructive and destructive [42] and need to be effectively managed instead of completely resolved, suggesting that communities should keep conflicts at a certain level to minimize the negative effects and enhance the positive effects, like satisfying the needs and expectations of the stakeholders [122]. There are five styles of handling interpersonal conflicts [121] with two dimensions [145] in the organizational context: (a) integrating (high concern for self and the other); (b) dominating (high concern for self and low concern for the other); (c) obliging (low concern for self and high concern for the other); (d) avoiding (low concern on both dimensions); and (e) compromising (middle on both dimensions).

Many scholars document the specific conflicts and management strategies in online communities. For example, the styles to manage task and relationship conflicts in open source development communities are using third-party intervention, coding in modularity, paralleling software development lines, and leaving the communities [151]. In virtual teams, members manage their conflict and negative emotion using third-party mediation, apology, explanation, positive reinforcement, and feedback-seeking behaviors [4]. Little work has directly applied the five styles in online communities. To our knowledge, Ishii's work is the first to directly apply these styles exploring online relationships [79]. Their work suggests that different computer-mediated communication technology (e-mail, text messaging vs. web camera) can influence users' perception of management styles and encourages exploration of a broader range of online communities. In line with their work, we directly apply Rahim's five management styles into live streaming communities and ask the following questions:

- **RQ1a:** How are perceived intragroup conflicts in the moderation team associated with the individual conflict management styles?
- **RQ1b:** What are the specific incidents of conflict, and how do mods handle them?

Online users apply cooperative management styles in their close relationships and avoid assertive styles if they want to continue the relationship [79]. Though relational explanation and encouragement cannot decrease the propensity to leave in goal-oriented communities because they fail to offer the insight to solve the problem and achieve the goal [77], it is still unclear in relationship-oriented communities how different community commitments influence the management styles. In addition, past research reported that text-based CMC diminishes status and power differences yet increases equality between communicators [137], and individuals can be aggressive toward one another [156]. Thus, anonymous users may take advantage of these characteristics and manage conflicts differently with someone they have never met. In live streaming communities, we asked the following research question:

- **RQ2:** How do moderators' commitments to the streamer and moderation experience influence conflict management styles?

7.3 Methods

This project was approved by the Institutional Review Board (IRB). We aimed to understand the relationships of mods' perceptions among conflicts, conflict management styles, and commitments to the streamer. We designed a survey to collect self-reported data from mods. At the beginning of the survey, we clarified that we were looking for content moderators in live streaming communities and this study would help us understand the conflict issue between mods and streamers. Participants had to consent to start the survey. The main survey includes three parts. The first part consists of general questions about their moderation experience, such as “how long have you been in the live streaming communities?” “How active do you moderate the chat?” and “what content is the channel focused on?” At the near end of the first part, we also had an open-ended question to ask mods to describe incidents of conflict with the streamer and how they handle it. This question can potentially help us gain context of conflict and conflict management. The second part includes the main variables about measurements of conflict, conflict management, and commitment. The third part includes demographic variables, such as age, race, gender, education, and country. The complete survey was in the supplementary file.

7.3.1 Participant Recruitment

We used a recruitment platform called Prolific[1] to collect the data. The platform used its user pool and automatically matched and distributed the survey to potential targets based on users’ self-reported information on its platform. About 500 people participated. We carefully set the survey and filtering questions to ensure the quality of the data. Specifically, we asked a multiple-choice question about their role in the live streaming community (Streamer/Broadcaster, Viewer/Normal user, Moderator (Mod), Other) at the beginning of the survey. Only participants who chose at least the “Moderator (Mod)” option were qualified for the study. This survey took about 10-15 minutes to complete. All responses with completion times that were less than 5 minutes were discarded. We monitored the survey progress and reviewed each participant’s completion in about a week. Each participant received the code to redeem $2 after the completion. In the middle of the survey, we also intentionally repeated a question as an attention-checking question. Participants should have the

[1] `https://prolific.co/`. Retrieved on March 14, 2022

same answer to prove they read the questions carefully. After we rejected and discarded responses through the filter question, attention-checking question, and completion time constraints, we finally had 240 qualified responses for analysis.

7.3.2 Participant Demographic

Among the 240 mods, 45.4% also identified them as viewer/normal users and 14.6% as a streamer. Participant's gender was 77.1% male, 22.1% female, 0.4% trans female, and 0.4% non-conforming. Participants were predominantly White (62.5%), followed by Hispanic/Latino (31.3%), Asian (5.0%), and African-American (3.8%); one participant preferred not to answer. Most participants had a bachelor's degree (29.6%), followed by graduated high school (27.1%), some college/no degree (26.3%), advanced degree (8.3%), associated degree (7.5%), and less than high school (1.3%). Most participants were young users: 18 to 24 years old (57.9%), 25 to 34 years old (31.3%), 35 to 44 years old (8.3%), and 45 to 55 years old (2.5%).

7.3.3 Survey Measure

The following items were initially developed in the organizational context and adapted to live streaming communities; some items that applied to the physical context were removed.

Task, Relationship, Process, and Normative Conflict in the Moderation Team We used Jehn's [80] eight-item scales to measure task conflict (M= 2.24, SD= .63, α= .75) and relationship conflict (M= 1.95, SD= .71, α= .85) in the moderation team. We used Jehn and Mannix's three-item [82] to measure the moderation process conflict in the moderation team (M= 2.00, SD= .76, α= .79). Responses were made on a 5-point scale where 1= "Never" and 5= "Always". We measured normative conflict (M= 2.38, SD= .80, α= .85) using Dahling and Gutworth's eight-item scales

[36]. Responses were made on a 5-point scale where 1= “Strongly Disagree” and 5= “Strongly Agree.”

Conflict-management Styles We adapted from Rahim’s 28-item conflict-management scales [121] to measure the five conflict-management styles (1= “Strongly Disagree” to 5 = “Strongly Agree”): integrating (M= 4.07 , SD= .57, α= .84), avoiding (M= 3.31, SD= .83, α= .79), dominating (M= 2.99, SD= .80, α= .79), obliging (M= 3.79, SD= .61, α= .85), compromising (M= 3.71, SD= .59, α= .70).

Commitment to the Streamer We measured commitment using the scales originally developed by Meyer and Allen [1] and adapted by Bateman et al. to online communities (1= “Strongly Disagree” to 5= “Strongly Agree”) [6]. They were continuance commitment to the streamer (CCtS) (M= 3.00, SD= .68 , α= .60), normative commitment to the streamer (NCtS) (M= 3.34, SD= .81, α= .80), affective commitment to the streamer (ACtS) (M= 3.89, SD= .71, α= .86).

7.3.4 Open-ended Question Analysis

The deductive content analysis aims to test previous theories, categories, and models in a different situation [48]. We followed a deductive approach to have the four types of conflict and five management styles as the structured categorization matrix. First, two authors went through the responses to prepare to code and decided to treat each response as a unit since most responses were short. Next, the leading author imported all data into ATLAS.ti [2] to iteratively code all responses in approximate three weeks with weekly calibration meetings with the second author to present quotes and discuss the fit. For example, codes such as *“streamer considers joke, but mods consider offensive”* and *“streamer wants to ban someone not violating rules”* formed a subcategory called “ discrepancy about rules” under normative conflict. Some

[2] https://atlasti.com/

responses about simply helping streamers or blocking viewers were not considered conflicts in the moderation team and were put aside. Some long quotes were coded with both conflict and management styles. We reported data roughly following the emphasis of the quote. If the description was detailed about conflict, we reported it under the conflict category; if the description was detailed about management, we reported it under management style.

7.4 Results

7.4.1 Descriptive Results

Most mods only moderated for one live streaming platform (77.5%), 20.0% did two to three platforms, and only 2.5% did more than three platforms. On the main platform they moderated, most of them moderated one channel (60.4%), and then 2-3 channels (32.1%), 4-5 channels (4.2%), 6-7 channels (2.5%), more than 7 channels (0.8%). The length being mods was in 1-2 years (40.4%), and then less than 1 year (33.8%), 2-3 years (12.1%), more than 4 years (7.9%), 3-4 years (5.8%). Regarding length staying in live streaming communities in general (watching, streaming, or moderating). Most mods used live streaming services for more than 4 years (31.3%), followed by 1-2 years (21.7%), less than 1 year (17.5%), 2-3 years (17.1%), and 3-4 years (12.5%). Most mods spent less than 12 hours in a week on moderation (50.4%), and then 12-24 hours (32.5%), 24- 36 hours (11.3%), 36-48 hours (3.8%), more than 48 hours (2.1%). Most mods were somewhat active to interact with viewers (54.2%), followed by very active (27.1%), not very active (17.9%), never (.0.8%). Similarly, Most mods were somewhat active to moderate the chat (48.3%), followed by very active (35.4%), not very active (14.2%), never (2.1%). The streaming content that they mainly moderated for was gaming (77.5%), followed by just chat (33.8%), art and music (14.2%), food and eating (7.9%), outdoor activity (6.3%), shopping (2.5%), others (8.8%) such as 3D modeling, technology, talk shows, education, sports and so forth.

Table 7.1 H1,2,3: Correlations among Conflicts and Commitments to the Streamer

Variables	1	2	3	4	5	6	7
1. Task conflict (Team)	1.00						
2. Relationship conflict (Team)	.73**	1.00					
3. Process conflict (Team)	.67**	.65**	1.00				
4. Normative conflict (Team)	.42**	.51**	.56**	1.00			
5. ACtS	-.01	-.07	-.02	-.13*	1.00		
6. CCtS	.02	.02	.07	.01	.29**	1.00	
7. NCtS	.16*	.10	.14*	.15*	.44**	.11	1.00

Note: [*] $p< .05$; [**] $p< .01$; [***] $p< .001$; $N= 240$; ACtS = Affective Commitment to the Streamer; CCtS = Continuance Commitment to the Streamer; NCtS = Normative Commitment to the Streamer.

7.4.2 Hypotheses Test

A Pearson's correlation analysis in Table 7.1 showed that relationship conflict in the moderation team was not associated with the ACtS ($r= -.07$, $p= .258$), and CCtS ($r= .02$, $p= .743$). Thus, **H1a and H1b were not supported**. Task conflict ($r= -.01$, $n= 240$, $p= .869$) and process conflict ($r= -.02$, $p= .789$) in the moderation team were not associated with ACtS. Thus, **H2a was not supported**. Task conflict ($r= .02$, $p= .706$) and process conflict ($r= .07$, $p= .259$) in the moderation team were not associated with CCtS. Thus, **H2b was not supported**. Normative conflict in the moderation team was positively associated with NCtS ($r= .15$, $p= .024$). Thus, **H3 was supported**.

Table 7.2 RQ1a: Correlations among Conflicts and Management Styles

Variables	1	2	3	4	5	6	7	8	9
1. Task conflict (Team)	1.00								
2. Relationship conflict (Team)	.73**	1.00							
3. Process conflict (Team)	.67**	.65**	1.00						
4. Normative conflict (Team)	.42**	.51**	.56**	1.00					
5. Integrating	-.08	-.15*	-.15*	-.24***	1.00				
6. Avoiding	.07	.05	.20**	.18**	.04	1.00			
7. Dominating	.24***	.25***	.22**	.23***	-.04	.05	1.00		
8. Obliging	-.14*	-.13*	-.04	-.08	.30**	.19**	.04	1.00	
9. Compromising	-.01	-.02	.04	-.03	.42**	.17**	.09	.12	1.00

Note: [*] p< .05; [**] p< .01; [***] p< .001; *N*= 240.

7.4.3 RQ1a: Relationship Between Conflict Types and Conflict Management Styles

A Pearson's correlation analysis in Table 7.2 showed that integrating was negatively associated with relationship (r= -.15, p= .023), process (r= -.15, p= .021), and normative conflict (r= -.24, p< .001) in the moderation team. Avoiding was positively associated with process (r= .20, p= .002) and normative conflict (r= .18, p= .006) in the moderation team. Dominating was positively associated with all four conflicts (task, r= .24, p< .001; relationship, r= .25, p< .001; process, r= .22, p= .001; normative, r= .23, p< .001) in the moderation team. Obliging was negatively associated with task (r= -.14, p= .032) and relationship (r= -.13, p= .045) conflict in the moderation team. Compromising is not associated with any type of conflict.

7.4.4 RQ1b: Incidents of Conflict and Conflict Management Styles

Most mods (about 68%) clearly expressed certain levels of conflict with the streamer. A group of mods expressed no specific conflict with the streamer or explained that they punished viewers (30.8%). Some mods mentioned the conflict between streamers in different channels or helping streamers with technical issues (about 1%). These are not considered intragroup conflicts between mods and streamers. We only reported the conflict between mods and streamers. In the following section, each quote, either short or long, represented one mod's opinion.

Normative Conflict Normative conflict in the channel can be separated into two subcategories: streamer's violation and discrepancy about rules (whether the comments/post should be considered a violation).

Streamer Violation Some streamers did *"not fully understand the rules"* or went off-topic and started doing *"something against terms of service."* Mods would remind streamers to adhere to the rules and help them to avoid norm-violation against the community guideline. Usually, streamers took their advice, and *"they are good about getting back on the topic."* One mod said that the streamer *"unknowingly did not follow up some rules regarding copyrighted content (mostly music tracks) but we got hold of the situation promptly, and the problem got solved smoothly."* Similarly, the streamer presented a bad act in the stream without notification. Mods sometimes even *"spam"* and *"annoy him "* to remind the streamer as a way to protect the streamer and the community, like this mod said, *"The streamer accidentally showed a bad word that is bannable on stream and didn't notice, so as the mods we had to spam him and annoy him hard so he would take down the stream and delete the VOD. In my opinion, the faster, the better, otherwise they'd get banned."*

Generally, when mods experienced normative conflict about the streamer's violation, they showed strong concerns to the streamer and would like to communicate with the streamer to remedy the behaviors, a typical integrating style. In rare cases, if the streamer insisted on not violating the rules or not listening to mod's suggestions, mods might quit and leave the community as an avoiding style. One mod said, *"The incident involved a streamer who kept making racist and offensive comments in a row.*

He later explained that it was a joke, but I was not okay with that, so I quit." Quitting moderation was an extreme case as a way to avoid conflict with the streamer.

Discrepancy About Rules The streamer and mods sometimes had a discrepant view about whether the content is offensive or not, such as *"difference in opinion regarding potential spam message"* and *"discrepancies to what could and couldn't be said in the live chat."* Sometimes, the mod considered it was offensive, but the streamer did not.

> Most recently a conflict of opinions happened, and that is what happens the most, even tho we mods work to keep things in order, the stream owner has his own idea of how he wants things to be, what he tolerates, and what he doesn't. When we end up disagreeing there's the problem, this time was about what a user in chat wrote and was actually someone he met playing a friend, so for me, it was offensive even if said in a jokingly way, but for the streamer, it was okay because It was said as a joke plus he was his friend. Basically, it got sorted out by talking and discussing.

According to this mod, the mod and the streamer finally reached an agreement and *"sorted out"* the conflict after discussion, though we did not know whether it was a punishment or permission. Sometimes, the streamer considered it was a violation, but the mods did not think so. One mod said, *"The streamer insists that I ban all the viewers who spam, but I believe that sometimes this can attract even more viewers and make the channel more alive. Of course, I don't mean spamming inappropriate things, but I mean spamming things related to the game the streamer plays."* The mod considered that spamming related to the streaming topic attracted viewers while the streamer did not allow any spam. Though mods provide suggestions and even argue with the moderation team, the streamer listened but might *"insist"* their attitudes toward the punishment.

> A user shared content in a specific channel, and the streamer (owner of the channel) asked me to remove the content due to being in a 'wrong channel'. I did not agree since what that user posted could be useful for many people who used that channel and declined it. The way I handle it was to give my opinion about the content but, either way, it was removed by other moderators not long after.

The mod usually handled this type of normative conflict by giving opinions. If the streamer accepted the advice after the discussion, it was an integrating style. If the streamer or other mods insisted on their attitudes, the mod had to compromise and accept the team's decision, which was a compromising style.

Process Conflict Some mods reported issues about communication, task assignment, and responsibility. They explained issues, suggested alternatives, or apologized if they made mistakes during the process. A few mods expressed the overload of the work due to the lack of enough mods. One mod said, *"They wanted me to be more proactive with their viewers and answer to every comment, which isn't possible taking into consideration that there are a lot of comments per stream, so we came to an agreement of what was expected of me during the streams."* Additionally, mods would like to discuss with the streamer how to handle the process conflict like hiring more mods to distribute tasks. Streamers considered these were good ideas and would implement them: *"I explained to the streamer that there are too few moderators for such a large group of recipients. He claimed that everything was fine, but in the end, he saw for himself that there were too few of us for such a large audience. I managed to convince him to find someone to help. Now he says it was a very good idea."* In these cases, mods handled conflicts by explaining what they did and suggesting what the streamer could do to reduce workload, a typical integrating style.

Sometimes, the task and responsibility were hard to meet the needs of both parties. The streamer and mods had to make a compromise. For example, one mod said, *"One conflict that comes to mind is that there are times I've been busy and was unable to moderate during the entire live stream, so the streamer had to moderate the chat himself. I just apologized, the streamer understood, and I moderated normally."* According to this mod, the streamer had high expectations beyond the mod's capability. The mod also admitted what he could do and apologized. The streamer accepted the fact and let the mod keep doing what he could do.

A few mods had different opinions about the streamer's performance and preference during the streaming process and would like to suggest the streamer behave in a certain way to facilitate community growth. One mod said, *"The streamer wanted to change the chat to subscriber-only mode, and I wanted to keep it public. I told him that keeping the chat public would increase his viewers, and he kept the chat public."* In this case, the streamer considered this good advice and took it. However, the streamer can also ignore their suggestion and leave it in the air, like this mod:*"We did get into an argument once because I told him he should use a microphone and a webcam so more people would join, and he didn't want to. It wasn't a heated argument, so it kind of blew off on itself."*

When experiencing process conflict, the mods would like to discuss and coordinate with the streamer no matter whether they finally reached an agreement, an integrating style. Sometimes, they had to make a compromise to consider the situation of both parties, a compromising style.

Relationship Conflict It is related to emotional and personal battlement with the streamer. Mods reported apparent relationship conflict with the streamer. The tension was usually caused by mistakenly blocking streamers' friends. For example, *"Streamer's friend started to insult him for jokes. I banned him because it was against* *the rules; I didn't know that was his friend, and streamer was angry on me."* Mods were at risk of losing mod status if they had a relationship conflict with the streamer. Sometimes, the streamer warned the mod to lose status, but the mod argued back: *"There was one person who broke like 5 rules so I timed out him for 10min, later on, the streamer messaged me to unban him because he was his friend and I had a choice to unban him or get kicked out of the mods team. I had an argument with him after a stream, but everything was fine after all."*

Communication or personally and gently handling the emotional streamer helped resolve the conflict. One mod said that he accidentally banned the streamer's close friends, making the streamer very angry and cancel his mod status, but a few days later, the streamer gave the mod a status again. The mods did not argue the issue with the streamer and fortunately got the status back. Alternatively, they might also talk with the streamer: *"The streamer started acting weird with me, he removed my mod, but after we talked, I got my mod back. I guess he was in a bad day."* If the discussion failed to reach an agreement, the mod might not *"continue the conversation,"* like this mod said, *"During the conversation about the election we did not agree in the podium, there was an emotional discussion with the use of bad words, to end it I just did not continue the conversation."* In this case, the mod tried an integrating style first and used an avoiding style if the former one didn't work.

Task Conflict It is the disagreement about the moderation action. Many mods reported the conflict regarding the punishment they should give to the violator. About 12 mods said that the streamer complained about *"being too strict in banning users for inappropriate comments"* (e.g.,. *"I was too strict with moderating the use of some emoticons"*, *"I was too hard on the banishing of people"*). Though the viewer violated the rules, streamers were very *"soft"* to some matters, but mods considered severe punishments.

> It happens more often than not that some viewers do not follow the rules (no ads, no caps, no asking for subs...), and as a consequence got banned. In those situations, we (= mods) just ban or timeout them for a while, and sometimes the streamers consider that we have been too strict (even though rules are rules and should be respected).

In the above case, both the streamer and mods mutually agreed that some viewers violated the rules, but the streamer considered mods' punishment such as ban and timeout to be too strict. Perhaps the streamer thought that frequent blocking hindered the viewership and was harmful to the micro community. However, mods might have different values.

> As a rule, we don't allow racial slurs in chat, in any context whatsoever. There's a lot of popular memes that involve the use of racial slurs and they get posted in the chat by viewers. Recently the streamer has asked me to ignore these racist memes, but I keep enforcing the rules, banning potential newcomers/subs. He thought this affected his subscription income, but I don't think we should allow this just because of the money.

This mod felt that the *"racial slurs"* should be banned while the streamer permitted the violation with the concern of losing subscription income. The mod used the authority to keep enforcing the rules and not taking the streamer's advice to *"ignore these racist memes,"* a dominating style.

Oppositely, the streamer sometimes required the mods to enforce the rules and actively moderate the chat while the mods had different opinions about punishment. One mod and the streamer showed different attitudes and punishments toward a troll comment: *"The streamer thought it was not OK while I thought it wasn't even worth it to give attention to a troll comment. I simply muted the viewer while the streamer wanted to give him an opportunity to discuss."* According to this mod, the normative conflict (whether troll was a violation) caused a task conflict (whether it should be blocked). Sometimes, the streamer might find the mod's opinion valuable after insisting on their opinions: *"He told me that I was too permissive with the chat and that then he could create a problem if his community got out of control... Later, as soon as I acted more harshly, it fell apart as several users complained about it. In the end, we solved it by talking, he defends more the attitude of his moderators since then."* In this case, the mod followed the streamer's suggestions and *"acted more harshly,"* but caused complaints. The mod had more expertise and experience about what was permissive or not and won the streamer's attitude.

Task Conflict and Integrating Mods were highly active in engaging and providing opinions to reach an agreement that satisfies both the streamer and them. For task conflict, they would usually either talk to the streamer to reach an agreement together or convince the streamer to allow or block viewers to support the micro community. One mod said, *"We have had personal disputes over certain toxic messages which we thought should have been banned or not. But nothing too heated, we discussed it over DMs and came to a mutual agreement."*

Several mods convinced the streamer by explaining and showing concerns to the streamer and the micro community. *"The streamer wanted me to ban people he didn't like personally, but who didn't break chat rules. I talked to him in private chat and convinced him it wasn't a good idea long term. We try to preach free speech."* This mod and streamer agreed that this was not a violation, but the streamer personally wanted to ban the viewer, the mods adhered to the rules and convinced the streamer not to do so. Similarly, another mod said, *"We had a discussion about if we needed to block people that are always being mean to others, we talk a lot, and I convinced him that the best thing for the rest of the community was to ban them."*

Though sometimes the violator was the streamer's friend, the mod would like to argue with the streamer, showing concern for the rest of the community and convincing the streamer to ask their friends to stop breaking the rules.

> We got into a conflict because some of his friends were spamming the chat (like in a joke or just messing around) and I wanted to ban them at least for the rest of the stream because they were making the chat unbearable for other users. The conflict was that he didn't want to ban them because he believed that was too much, but I tried to argue that they were affecting other members in the chat that are more important because honestly his friends were still going to continue be liking and commenting on posts but other people could go. He told me he would talk with them, he did and after a couple of minutes, the spamming stopped.

In this case, the mod didn't ban the violator because the streamer *"believed that was too much,"* but the streamer took the mod's advice and asked the violator to stop the violation in the chat.

Task Conflict and Obliging About 13 mods explicitly reported that they *"stopped arguing and gave in"* if they had a task conflict, such as reversing punishment and following the order to punish someone, though they disagreed about the punishment. For example, *"I kicked a user out of chat that I felt violated the streamer's rules but they wanted them to stay. It's their channel so I brought them back,"* said one mod. Sometimes, the streamer considered political and controversial themes violations and asked mods to make severe punishments. One mod said, *"He asked me to ban everyone that remotely mentioned politics, I thought it was a bit harsh, but I still did it."* In this case, though the mod felt that the punishment was harsh but still followed the order. Conversely, the mod sometimes considered a harsh punishment but the streamer didn't feel so.

> I banned a user for saying something, which I deemed was offensive to the streamer and in general, but the streamer didn't agree with me. He didn't think it was worthy of a permanent ban and wanted me to change it to a temporary one. I eventually did what the streamer asked, but I strongly disagreed that what the user said was acceptable.

According to this mod, the disagreement was between a *"permanent ban"* or a *"temporary one."* Though the mod *"strongly disagreed"* with the streamer, the mod eventually followed the streamer's order and changed the punishment. If mods didn't follow the streamer's order, they might lose their mod status so they had to oblige, as this mod said, *"There was a certain occasion of a troll in the comments cursing on the streamer. I offered to ban the troll, but the streamer wanted to get in conflict with him exchanging curses live because it was more fun—thus canceling my purpose as a mod. I had to oblige."*

7.4.5 RQ2: Commitment to the Streamer and Conflict Management Styles

We ran a series of linear regression models with mod's commitment to the streamer and moderation experience as independent variables and five conflict management styles as dependent variables (see Table 7.3). For integrating, the model explained 5% variance, *adjust* R^2= .05, *F(8,231)*= 2.70, *p*= .007. Only ACtS was positively related to it. For avoiding, the model explained 7% variance, *adjust* R^2= .07, *F(8,231)*= 3.13, *p*= .002. Both ACtS and hours of moderation weekly are positively related to it. For Obliging, the model explained 19% variance, *adjust* R^2= .19, *F(8,231)*= 7.86, *p*< .001. Both NCtS and ACtS are positively related to obliging style; additionally, length in the community is positively related, but the length of being mod is negatively related to it. For dominating, the model is not significant (*adjust* R^2= .02, *F(8,231)*= 1.68, *p*= .104), though length of being mod is positively related to it. For compromising,

Table 7.3 RQ2: Regression Model Examining the Effect of Commitments to the Streamer on Conflict Management Styles

Variables	Integrating	Avoiding	Dominating	Obliging	Compromising
Commitments					
CCtS	.01	.06	.08	.08	-.02
NCtS	.07	.25***	.10	.23**	.09
ACtS	.19*	-.09	.02	.23**	.03
Moderation experience					
Length in the community	-.08	-.12	-.09	.18**	-.18*
Length of being mod	.15	-.02	.17*	-.16*	.05
Hours of moderation weekly	-.02	.15*	-.04	-.02	.07
Active interacting	.07	.12	.06	.00	.01
Active modding	.05	-.13	.05	.06	.07
Adjust R^2	.05	.07	.02	.19	.01
F	2.70**	3.13**	1.68	7.86***	1.39

Note: [*] p<.05; [**] p<.01; [***] p<.001; all β values are standardized coefficients; ACtS = Affective Commitment to the Streamer; CCtS = Continuance Commitment to the Streamer; NCtS = Normative Commitment to the Streamer.

the model is not significant (*adjust* R^2= .01, *F(8,231)*= 1.39, *p*= .202), though length in the community is negatively related to it.

7.5 Discussion

This research explores the relationships among mods' intragroup conflict, conflict management styles, and their commitments to the streamer in live streaming communities. The findings provide a nuanced understanding of the conflict in the community moderation team and can be potentially generalized to live streaming communities or new forms of media or other platforms applying a similar governance structure. This research also can potentially foster productive relationships between community mods and admins and help them build effective moderation teams.

7.5.1 Conflicts in the Team and Commitments to the Streamer Are Independent

The first part of the research explores the relationship between conflicts and commitments. Generally, conflicts in the moderation team are independent of mods' commitments to the streamer. This stands in contrast to research showing significant relationships between conflicts and commitments. Speculatively, some research has specified the attribution of commitment. For example, team members show greater commitment to the decision if they perceive the decision process as fairer [98]. Task conflict stimulates members' commitment to the task if team members' voices and ideas are fairly incorporated into the group decisions [8]. In this study, we set the commitment attribution to the streamer, but the moderation tasks are toward the viewer/violators. Mods realize the conflict in the team had nothing to do with their emotional attachment to the streamer and their intention to stay with the streamer. In this sense, we provide a nuanced understanding of the relationship between conflicts and commitments in community moderation. Future work can investigate (1) how conflict might affect mods' commitment to other entities such as viewers, other mods, or even task itself, (2) the impact of intragroup conflict on third-party stakeholders and the roles they play in the conflict-commitment relationship in online communities in general.

The positive but weak association between normative conflict and normative commitment to the streamer suggests that when normative conflict increases in the team, mods have a stronger sense of obligation to support the streamer, and vice versa. Prior work suggests that members with high normative commitments in the online community are more likely to engage in constructive behaviors that preserve the community [6]. Mods are dedicated viewers or streamers' friends sharing the common value about the micro community [157, 158]. When mods experience normative conflict with the streamer, they offer support to the streamer regarding the streamer's violation and the discrepancy of rules.

The positive relationship raises questions about whether the micro community leader/streamer should increase the normative conflict if it is beneficial to keep the mod's normative commitment to the streamer. The average score of normative conflict in the moderation team shows that the conflict level is relatively low in live streaming communities. Prior work suggests that conflict should be effectively managed at the individual and group levels [81], and enforcing a particular rule with much normative conflict increases the possibility to lead to counter-punishment and even feud [117]. How to balance the amount of normative conflict and its effectiveness to influence mod's normative commitment to the streamer needs further investigation.

7.5.2 Active and Cooperative Style Versus Passive and Assertive Style

The second part of the research explores the relationship between conflict and conflict management styles. To cooperatively manage conflicts, team members tend to use conflict to promote compatible goals and resolve them with integrating and high-quality solutions for mutual benefit; consequently, cooperativeness can increase procedural justice and lead to team innovation [147]. We found that, generally, when mods use dominating and avoiding styles to handle the conflicts with the streamer, the conflicts in the moderation team are likely to increase; when mods use integrating and obliging styles to manage conflicts with the streamer, the conflicts are likely to decrease, suggesting that, for the individual mod, active and cooperative styles to handle conflict with the streamer can be potentially more effective than passive and assertive styles. Cooperative styles like integrating to manage conflicts can also increase their perceptions of interpersonal outcomes, such as belonging and appreciation for others [154].

However, conflict in teams is a complex and dynamic process changing over time and impacted by many factors [76], and the management styles are also highly contingent; no one best approach can deal with different situations effectively [122].During the conflict management process [123], the real-time nature of live streaming requires mods to identify violations and make quick decisions. The problem-solving process causes conflict in the team. As conflicts arise and evolve, mods manage different conflicts with contingent styles. Assertive styles can also be be effective but is highly contingent on individual and collaborative factors such as the number of mods, the credibility of mods, and the overall opinion valence in the team [70]. Similarly, we found that moderation experience affects management styles (e.g., mods with higher tenure of the community are more likely to use obliging, but experienced mods are more likely to use dominating), but we don't know whether these styles are effective. Additionally, how to balance cooperative and assertive styles and increase the effectiveness of conflict management overall also needs future investigation.

Such results show that, though mods can actively propose and argue with the streamer, the streamer is the core in the hierarchy, indicating that the communication among mods and the streamer is not exactly democratic. We saw that mods' autonomy in live streaming communities is somewhat restricted, compared with mods making decisions on other online communities. The qualitative results show that mods can use dominating style to handle task conflict (e.g., keep enforcing rules and ban racial slurs instead of taking the streamer's advice to ignore them). However, we don't know what happened next. Streamers can accept mods' actions and move on or insist on their opinions and cause more task conflict, even transfer the task conflict into a relationship conflict to risk losing mod status. Mods have to use either obliging or avoid in the end. It seems like mods are forced to be cooperative, to some extent. Further research may examine the power structure between streamers and mods and explore how these power dynamics influence conflict dynamics and conflict management styles.

7.5.3 High Concerns for the Streamer or Low Concerns for Mods Themselves

The last part of the research explores the relationship between commitment and conflict management styles. In general, mods with strong commitments to the streamer would like to apply styles that show either high concerns for the streamer or low concerns for themselves. In line with prior work indicating that different commitments affect different kinds of online behaviors [6], we contribute to a nuanced understanding of how different commitments to the streamer predicting their conflict management styles in live streaming moderation teams. Prior work suggests that users who are in a close relationship or intend to build a close relationship with others will use the integrating and obliging style [79]. Similarly, we found that mods having a stronger affective commitment to the streamer are more likely to show high concerns for the streamer and use integrating and obliging styles, suggesting that if mods have a high emotional attachment to the streamer, they are more likely to provide their opinions and discuss with the streamer; they either finally reach an agreement or follow the streamer's order.

Additionally, mods with a stronger normative commitment to the streamer are more likely to show low concerns for themselves and use avoiding and obliging styles, suggesting that if mods morally feel they should help and support the streamer, they would like to either handle it personally with salience or provide suggestions to show care more about the streamer. It seems like mods with strong commitments to the streamer (either affective or normative) would ultimately try to satisfy the streamer's needs. However, as we showed in the aforementioned section, the management styles are contingent. The static regression analysis can't reflect the dynamic of management styles in the moderation team.

7.5.4 Design Implications

Clarify Norms and Punishment to Avoid Too Much Task and Normative Conflict in the Team The prominent category about task conflict under RQ2 suggests that though mods and the streamer agree about the violation, which is clearly stated in the chat rule or channel rule but have different attitudes toward a punishment to the violator in many cases. Many community rules use prescriptive and restrictive norms to show what is allowed or not [49], but rarely specific the consequence. As a way to avoid task conflict, the rule statements should indicate the consequence of the violation. However, too much transparency can also cause trouble and allow violators to strategically game the moderation system [43]. There is a need to balance effectiveness with fairness and transparency in the moderation mechanisms [129]. Research has shown that the mods' setting and view are different from the viewer's view and that mods can have access to a lot of information invisible to the public [19]. We propose an alternative mechanism to show clear rules with consequences in different scenarios, which is only visible to the moderation team but invisible to the public. For example, on the live streaming platform Twitch, designers can develop a two-layer chat rule with a switch button from the mod's view, mods can easily switch between the general rule display and the more specific rule display with decision suggestions. The public chat rule focuses on the clarity of the rule, while the private rules focus more on the consequence of each scenario.

Balance the Power Dynamics to Provide More Support to Mods in the Micro Community Live streaming platforms allow streamers to easily grant or revoke privileges from mods, but it is still unclear whether the design of the platform encourages streamers to be in charge, or mods would ultimately default to the streamer regardless because the micro community is centered around the streamer. Possibly, mods' deference to the streamer is a result of the system structure because mods responded to the mod status loss with passive styles from avoiding, compromising, and obliging. We propose a mechanism that facilitates the mod's appealing process or increases the streamer's barrier to arbitrarily cancel or entitle the mod's status. For example, on Twitch, the designer can consider adding a two-side agreement mechanism (e.g., a pop-up window to ask mods and streamers to agree to the terms of service), after both the streamer and the mod agree to entitle mod status or revoke. Additionally, it can also open a specific channel to hear both the streamer's and mods' voices and handle the streamer-mods conflict when they encounter trouble during the entitlement or revocation process. We don't know how it will affect streamers' thoughts about mod selection; maybe it will demotivate streamers to select mods or increase the conflict with mods since they have more power in the hierarchy. Understanding the ways in which we balance the support to the mods and the protection to the streamer's benefits should be further investigated.

7.5.5 Limitations and Future Work

There are several limitations to this study. We asked for general opinions of mods in live streaming communities without specifying the platforms. Prior work shows that the history, policy, or culture of the platforms might also influence mod's perceived roles and responsibilities [130], indicating the potential difference between conflict management styles and their relationships with types of conflict and commitment. Future work can consider different platforms to enrich the understanding of relationships in this study. Second, the conflict in the moderation team includes conflict with both mods and the streamer, and the styles in this study are mainly about conflict management with the streamer. Though we know that the streamer has the authority to make the final decision, and mods discuss with the streamer when they can not make a decision [19], we still don't know how mods handle the conflict with other mods, and how much conflict among mods. Third, we only considered the conflict and management styles from mods' view. We don't know how streamers as team leaders perceive the conflict and whether they would apply different styles. Prior work suggests subordinates using a obliging style with supervisors experiences more interpersonal conflict, but supervisors suing a integrating style with subordinates experiences more interpersonal conflict as well [154]. Future work can investigate conflict management from streamers' perspective. Fourth, we don't consider antecedents of conflicts such as culture and language differences [5, 146, 61, 71] and other factors, such as informational, social, and value diversity, in the our analysis [83]. They can significantly affect different types of conflict, requiring future work. Lastly, we only asked conceptual questions about these measures to reflect mods' perceptions. Future work can consider collecting behavioral data like actual instances of conflict complied from a moderation log at scale to validate these findings.

7.6 Conclusion

In this work, we aimed to understand the triangular relationships (types of conflict, mods' conflict management styles, and mods' commitment to the streamer) in the moderation team in live streaming communities. Conflicts and commitments to the streamer are independent, though the normative conflict in the moderation team is weakly associated with mod's normative commitment to the streamer. In general, active and cooperative styles (obliging and integrating) can be more effective than passive and assertive styles (avoiding and dominating) to manage conflict in the moderation team. Mods who have a strong normative commitment to the streamer are more likely to use avoiding and obliging styles to handle conflict with the streamer, and mods who have a strong affective commitment to the streamer are more likely to use integrating and obliging styles to handle conflict with the streamer. Future research and implications are also discussed.

CHAPTER 8

CONCLUSION

To regulate the negative content and maintain the civic discourse, a massive workforce of people moderates behind the scenes with the assistance of algorithms and moderation tools. With the adoption of new technology and the evolvement of communities, the moderation practice faces new challenges. The hidden labor of volunteer moderators in live streaming communities face challenges caused by synchronicity and ephemerality. My dissertation focuses on the volunteer moderator in interactive media with real-time affordances. I aim to understand volunteer moderators' relationships with viewers and the streamer, identify the challenges they face during the moderation process, and recommend possible social and technical interventions to maintain a safe online space.

First, I interviewed 21 Twitch moderators and mapped the moderation strategies in live streaming communities, revealing the visible and performative work of volunteer moderators in real-time moderation, and discussing possible socio-technical interventions to reduce the information overload [19]. I also categorized current moderation tools and highlighted the opportunities for bot design in live streaming communities [15].

Second, I aimed to understand how volunteer moderators on Twitch created profiles of violators before they decided on what action they took with the violator and applied video observation and interview methods. I used criminal profiling as a lens [75]. I found that profiling improved moderators' understanding of violators, and they engaged in complex practices of evidence collection and documentation to create these profiles [18].

Last, As each micro-community attempts to set its guidelines, it is common for mods and the streamer to disagree on handling various situations. I applied a mixed-method with survey data collection and target moderators in live streaming communities to statistically test the relationships among the perceived conflict, moderators' commitment to the streamer, and the conflict management styles. This research explored the triangular relationships among conflict types in the moderation team, moderators' conflict management styles, and moderators' commitment to the streamer.

Through taking a three-phased approach to understand the moderators work with both the streamer and viewers in live streaming communities, I highlighted the moderation challenges caused by the affordances of new technology and showed moderators' understanding of and relationship with stakeholders in the community. This work showed the potential to guide community moderation and maintenance with socio-technical interventions in new forms of social media with high interactivity and synchronicity.

REFERENCES

[1] Natalie J. Allen and John P. Meyer. The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology*, 63(1):1–18, 3 1990.

[2] Allen C. Amason. Distinguishing the effects of functional and dysfunctional conflict on strategic decision making: Resolving a paradox for top management teams. *Academy of Management Journal*, 39(1):123–148, 1996.

[3] AnyKey. Barriers to Inclusion and Retention: The Role of Community Management and Moderation Whitepaper. Technical report, AnyKey, 4 2016.

[4] Oluremi B Ayoko, Alison M Konrad, and Maree V Boyle. Online work: Managing conflict and emotions for performance in virtual teams. *European Management Journal*, 30(2):156–174, 2012.

[5] Keivan Bahmani, Zhaleh Semnani-Azad, Wendi L. Adair, and Katia Sycara. Computer Mediated Communication in Negotiations: The Effect of Intragroup Faultlines on Intergroup Communication and Outcomes. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[6] Patrick J Bateman, Peter H Gray, and Brian S Butler. The impact of community commitment on participation in online communities. *Information Systems Research*, 22(4):841–854, 2011.

[7] Howard S. Becker. Notes on the Concept of Commitment. *American Journal of Sociology*, 66(1):32–40, 1960.

[8] Kristin J. Behfar, Elizabeth A. Mannix, Randall S. Peterson, and William M. Trochim. Conflict in small groups: The meaning and consequences of process conflict. *Small Group Research*, 42(2):127–176, 2011.

[9] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Proceedings of the 9th International Conference on Social Informatics*, pages 405–415. Springer International Publishing, 2017.

[10] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. When online harassment is perceived as justified. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, pages 22–31, 2018.

[11] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19, 12 2017.

[12] Lindsay Blackwell, Mark Handel, Sarah T. Roberts, Amy Bruckman, and Kimberly Voll. Understanding "Bad Actors" Online. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*, pages 1–7, 2018.

[13] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 1 2006.

[14] Jie Cai, Cameron Guanlao, and Donghee Y. Wohn. Understanding Rules in Live Streaming Micro Communities on Twitch. In *Proceedings of the ACM International Conference on Interactive Media Experiences*, pages 290–295, 2021.

[15] Jie Cai and Donghee Y. Wohn. Categorizing Live Streaming Moderation Tools. *International Journal of Interactive Communication Systems and Technologies*, 9(2):36–50, 7 2019.

[16] Jie Cai and Donghee Y. Wohn. Live Streaming Commerce: Uses and Gratifications Approach to Understanding Consumers' Motivations. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2548–2557, 2019.

[17] Jie Cai and Donghee Y Wohn. What are Effective Strategies of Handling Harassment on Twitch?: Users' Perspectives. In *Conference Companion Publication of the ACM on Computer Supported Cooperative Work and Social Computing*, pages 166–170, 2019.

[18] Jie Cai and Donghee Y. Wohn. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.

[19] Jie Cai, Donghee Y. Wohn, and Mashael Almoqbel. Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities. In *Proceedings of ACM International Conference on Interactive Media Experiences*, pages 61–72, 2021.

[20] Jie Cai, Donghee Y. Wohn, Ankit Mittal, and Dhanush Sureshbabu. Utilitarian and Hedonic Motivations for Live Streaming Shopping. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, pages 81–88, 2018.

[21] Kevin M. Carlsmith and John M. Darley. Psychological Aspects of Retributive Justice. In *Advances in Experimental Social Psychology*, volume 40, pages 193–236. Elsevier Inc., 2008.

[22] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. Norms Matter: Contrasting Social Support Around Behavior Change in Online Weight Loss Communities. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

[23] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, pages 1201–1213, 2 2016.

[24] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

[25] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017.

[26] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.

[27] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 3175–3187, 2017.

[28] Jonathan Chang and Cristian Danescu-Niculescu-Mizil. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *Proceedings of the World Wide Web Conference*, pages 184–195, 2019.

[29] Ed H Chi. Transient User Profiling. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–4, 2004.

[30] W. Jerry Chisum and Joseph M. Rynearson. *Evidence and Crime Scene Reconstruction*. National Crime Scene Investigation and Training, Redding, CA, 5th edition, 1997.

[31] W. Jerry Chisum and Brent E. Turvey. Methods of Crime Reconstruction. In *Crime Reconstruction*, chapter 8, pages 179–209. Academic Press, 2 edition, 2011.

[32] Janine Natalya Clark. The three Rs: retributive justice, restorative justice, and reconciliation. *Contemporary Justice Review*, 11(4):331–350, 2008.

[33] Maxime Clément and Matthieu J. Guitton. Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia. *Computers in Human Behavior*, 50(1):66–75, 2015.

[34] John Cook and Toby Wall. New work attitude measures of trust, organizational commitment and personal need non-fulfilment. *Journal of Occupational Psychology*, 53(1):39–52, 3 1980.

[35] Kate Crawford and Tarleton Gillespie. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428, 3 2016.

[36] Jason J. Dahling and Melissa B. Gutworth. Loyal rebels? A test of the normative conflict model of constructive deviance. *Journal of Organizational Behavior*, 38(8):1167–1182, 10 2017.

[37] Link Daniel, Hellingrath Bernd, and De Groeve Tom. Twitter integration and content moderation in GDACSmobile. In *Proceedings of the 10th International ISCRAM Conference*, pages 67–71, 2013.

[38] Paul B. de Laat. Profiling vandalism in Wikipedia: A Schauerian approach to justification. *Ethics and Information Technology*, 18(2):131–148, 6 2016.

[39] Michael A. De Vito, Darren Gergle, and Jeremy Birnholtz. "Algorithms ruin everything": #RIPTwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 3163–3174, 2017.

[40] Frank R C De Wit, Lindred L Greer, and Karen A Jehn. Supplemental Material for The Paradox of Intragroup Conflict: A Meta-Analysis. *Journal of Applied Psychology*, 97(2):360–390, 2012.

[41] Leslie A. DeChurch and Michelle A. Marks. Maximizing the benefits of task conflict: The role of conflict management. *International Journal of Conflict Management*, 12(1):4–22, 2001.

[42] Morton Deutsch. The Resolution of Conflict: Constructive and Destructive Processes. *American Behavioral Scientist*, 17(2):248–248, 1973.

[43] Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016.

[44] Bryan Dosono and Bryan Semaan. Moderation Practices as Emotional Labor in Sustaining Online Communities. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[45] Craig Dowden, Craig Bennell, and Sarah Bloomfield. Advances in Offender Profiling: A Systematic Review of the Profiling Literature Published Over the Past Three Decades. *Journal of Police and Criminal Psychology*, 22(1):44–56, 2007.

[46] Janna Lynn Dupree, Richard Devries, Daniel M. Berry, and Edward Lank. Privacy personas: Clustering users via attitudes and behaviors toward security practices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 5228–5239, 2016.

[47] Steven A. Egger. Psychological profiling: Past, Present, and Future. *Journal of Contemporary Criminal Justice*, 15(3):242–261, 1999.

[48] Satu Elo and Helvi Kyngäs. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115, 4 2008.

[49] Casey Fiesler, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R Brubaker. Reddit Rules! Characterizing an Ecosystem of Governance. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, pages 72–81, 2018.

[50] Anna Filippova and Hichang Cho. Mudslinging and manners: Unpacking conflict in free and open source software. In *Proceedings of the ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, pages 1393–1403, 2015.

[51] Andrew T Fiore, Lindsay Shaw Taylor, G.A. Mendelsohn, and Marti Hearst. Assessing attractiveness in online dating profiles. In *Proceeding of the ACM Conference on Human Factors in Computing Systems*, page 797–806, New York, New York, USA, 2008. ACM Press.

[52] Bryanna Fox and David P Farrington. What have we learned from offender profiling? A systematic review and meta-analysis of 40 years of research. *Psychological Bulletin*, 144(12):1247–1274, 2018.

[53] Mathilde B Friedländer. Streamer motives and user-generated content on social live-streaming services. *Journal of Information Science Theory and Practice*, 55(11):65–84, 2017.

[54] Vernon J. Geberth. Psychological Profiling. *Law and Order*, 29(9), 1981.

[55] Ysabel Gerrard. Beyond the hashtag: Circumventing content moderation on social media. *New Media and Society*, 20(12):4492–4511, 2018.

[56] Eric Gilbert. Widespread underprovision on Reddit. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 803–808, 2013.

[57] Tarleton Gillespie. The politics of 'platforms'. *New Media & Society*, 12(3):347–364, 5 2010.

[58] Tarleton Gillespie. Governance of and by platforms. In Jean Burgess, Thomas Poell, and Alice Marwick, editors, *SAGE Handbook of Social Media*. SAGE Publications Ltd, 2017.

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

[60] Debbie Ging and James O' Higgins Norman. Cyberbullying, conflict management or just messing? Teenage girls' understandings and experiences of gender, friendship, and conflict on Facebook in an Irish second-level school. *Feminist Media Studies*, 16(5):805–821, 2016.

[61] Carolina Gomez and Kimberly A Taylor. Cultural differences in conflict resolution strategies: A US–Mexico comparison. *International Journal of Cross Cultural Management*, 18(1):33–51, 4 2018.

[62] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society*, 7(1):1–15, 2020.

[63] James Grimmelmann. The Virtues of Moderation. *Yale Journal of Law and Technology*, 17(1):68, 2015.

[64] Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, and Dawei Yin. Hierarchical user profiling for e-commerce recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 223–231, 2020.

[65] Harold Guetzkow and John Gyr. An Analysis of Conflict in Decision-Making Groups. *Human Relations*, 7(3):367–382, 1954.

[66] Chloe Hadavas. The Future of Free Speech Online May Depend on This Database. *Slate*, 8 2020.

[67] Oliver L Haimson and John C Tang. What Makes Live Events Engaging on Facebook Live, Periscope, and Snapchat. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 48–60, 2017.

[68] William A Hamilton, Oliver Garretson, and Andruid Kerne. Streaming on Twitch: Fostering Participatory Communities of Play within Live Mixed Media. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1315–1324, 2014.

[69] Jeffrey T Hancock, Catalina Toma, and Nicole Ellison. The truth about lying in online dating profiles. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 449–452, 4 2007.

[70] Florian Hauser, Julia Hautz, Katja Hutter, and Johann Füller. Firestorms: Modeling conflict diffusion and management strategies in online communities. *Journal of Strategic Information Systems*, 26(4):285–321, 2017.

[71] Helen Ai He, Naomi Yamashita, Chat Wacharamanotham, Andrea B. Horn, Jenny Schmid, and Elaine M. Huang. Two Sides to Every Story: Mitigating Intercultural Conflict through Automated Feedback and Shared Self-Reflections in Global Virtual Teams. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–21, 12 2017.

[72] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: The dynamics of the "location" field in user profiles. In *Proceedings of ACM Conference on Human Factors in Computing Systems*, pages 237–246, 2011.

[73] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. Searching for safety online: Managing "trolling" in a feminist forum. *Information Society*, 18(5):371–384, 2002.

[74] Scotia J. Hicks and Bruce D. Sales. *Criminal profiling: Developing an effective science and practice.* American Psychological Association, Washington, 2006.

[75] Scotia J. Hicks and Bruce D. Sales. A Scientific Model of Profiling. In *Criminal profiling: Developing an effective science and practice.*, pages 207–230. American Psychological Association, 2007.

[76] Pamela J. Hinds and Diane E. Bailey. Out of Sight, Out of Sync: Understanding Conflict in Distributed Teams. *Organization Science*, 14(6):615–632, 2003.

[77] Wenjian Huang, Tun Lu, Haiyi Zhu, Guo Li, and Ning Gu. Effectiveness of Conflict Management Strategies in Peer Review Process of Online Collaboration Projects. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, pages 717–728, 2 2016.

[78] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–12, 4 2020.

[79] Kumi Ishii. Conflict management in online relationships. *Cyberpsychology, Behavior, and Social Networking*, 13(4):365–370, 2010.

[80] Karen A. Jehn. A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative Science Quarterly*, 40(2):256–282, 1995.

[81] Karen A Jehn and Corinne Bendersky. Intragroup Conflict in Organizations: A Contingency Perspective on the Conflict-Outcome Relationship. *Research in Organizational Behavior*, 25:187–242, 1 2003.

[82] Karen A Jehn and Elizabeth A Mannix. The dynamic nature of conflict: A longitudinal study of intragroup conflict and group performance. *Academy of Management Journal*, 44(2):238–251, 2001.

[83] Karen A. Jehn, Gregory B. Northcraft, and Margaret A. Neale. Why Differences Make a Difference: A Field Study of Diversity, Conflict and Performance in Workgroups. *Administrative Science Quarterly*, 44(4):741–763, 12 1999.

[84] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. “Did You Suspect the Post Would be Removed?”: Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019.

[85] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction*, 26(5):35, 2019.

[86] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019.

[87] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction*, 25(2):1–33, 2018.

[88] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. Moderation Challenges in Voice-based Online Communities on Discord. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 11 2019.

[89] Mladen Karan and Jaň Snajder. Preemptive Toxic Language Detection in Wikipedia Comments Using Thread-Level Context. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 129–134, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

[90] Charles Kiene, Kate Grandprey-Shores, Eshwar Chandrasekharan, Shagun Jhaver, Jialun ”Aaron” Jiang, Brianna Dym, Joseph Seering, Sarah Gilbert, Kat Lo, Donghee Yvette Wohn, and Bryan Dosono. Volunteer Work: Mapping the Future of Moderation Research. In *Conference Companion Publication of the ACM on Computer Supported Cooperative Work and Social Computing*, pages 492–497, 2019.

[91] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. Technological frames and user innovation: Exploring technological change in community moderation teams. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[92] Sara Kiesler, Robert E. Kraut, Paul Resnick, and Aniket Kittur. Regulating Behavior in Online Communities. In Robert E. Kraut and Paul Resnick, editors, *Building Successful Online Communities: Evidence-Based Social Design*, chapter 4, pages 125–177. The MIT Press, 2012.

[93] Aniket Kittur and Robert E. Kraut. Beyond Wikipedia: Coordination and Conflict in Online Production Groups. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, page 215–224, 2010.

[94] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 453–462, 4 2007.

[95] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131(6):1598–1670, 2018.

[96] Richard N Kocsis. Criminal Psychological Profiling: Validities and Abilities. *International Journal of Offender Therapy and Comparative Criminology*, 47(2):126–144, 2003.

[97] Richard N. Kocsis, Harvey J. Irwin, Andrew F. Hayes, and Ronald Nunn. Expertise in psychological profiling: A comparative assessment. *Journal of Interpersonal Violence*, 15(3):311–331, 2000.

[98] M. Audrey Korsgaard, David M. Schweiger, and Harry J. Sapienza. Building Commitment, Attachment, and Trust in Strategic Decision-Making Teams: The Role of Procedural Justice. *Academy of Management Journal*, 38(1):60–84, 1995.

[99] Cliff Lampe, Nicole Ellison, and Charles Steinfield. A familiar face(book): profile elements as signals in an online social network. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 435–444, 2007.

[100] Cliff Lampe and Paul Resnick. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 543–550, 2004.

[101] Anti-Defamation League. Online Hate and Harassment: The American Experience 2020. Technical report, Anti-Defamation League, New York, NY, USA, 6 2020.

[102] Xiaocen Liu. Live streaming in China: boom market, business model and risk regulation. *Journal of Residuals Science & Technology*, 13(8):1–7, 2016.

[103] Claudia Lo. *When All You Have is a Banhammer: The Social and Communicative Work of Volunteer Moderators.* PhD thesis, Massachusetts Insitute of Technology, 2018.

[104] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. “Could You Define That in Bot Terms”?: Requesting, Creating and Using Bots on Reddit. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 3488–3500, New York, NY, USA, 5 2017. ACM.

[105] Zhicong Lu, Michelle Annett, and Daniel Wigdor. Vicariously Experiencing it all without Going Outside: A Study of Outdoor Livestreaming in China. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28, 2019.

[106] Zhicong Lu, Haijun Xia, Seongkook Heo, and Daniel Wigdor. You Watch, You Give, and You Engage: A Study of Live Streaming Practices in China. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[107] Yannis Markovits, Ann J Davis, and Rolf Van Dick. Organizational commitment profiles and job satisfaction among Greek private and public sector employees. *International Journal of Cross Cultural Management*, 7(1):77–99, 2007.

[108] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. Impression Formation in Online Peer Production: Activity Traces and Personal Profiles in GitHub. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, pages 117–128, 2013.

[109] Luck Marthinusen. Social media trends in 2018: Live streaming dominates the social media landscape, 2017.

[110] J. Nathan Matias. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 5(2):12, 2019.

[111] Aiden McGillicuddy, Jean Grégoire Bernard, and Jocelyn Cranefield. Controlling bad behavior in online communities: An examination of moderation work. In *Proceeding of the 36th International Conference on Information Systems*, pages 1–11, 2016.

[112] Francisco J Medina, Lourdes Munduate, Miguel A Dorado, Inés Martínez, and José M Guerra. Types of intragroup conflict and affective reactions. *Journal of Managerial Psychology*, 20(3-4):219–230, 2005.

[113] Stuart E Middleton, Nigel R Shadbolt, and David C De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22(1):54–88, 2004.

[114] Claudia Müller-Birn, Leonhard Dobusch, and James D. Herbsleb. Work-to-rule: the emergence of algorithmic governance in Wikipedia. In *Proceedings of the 6th International Conference on Communities and Technologies*, pages 80–89, 2013.

[115] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383, 2018.

[116] NFSTC. A Simplified Guide to Crime Scene Investigation. Technical report, National Forensic Science Technology Center, 2013.

[117] Nikos Nikiforakis, Charles N. Noussair, and Tom Wilkening. Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics*, 96(9-10):797–807, 10 2012.

[118] Elinor Ostrom. A Behavioral Approach to the Rational Choice Theory of Collective Action. *American Political Science Review*, 92(1):1–22, 1998.

[119] Elinor Ostrom. *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press, Cambridge, 2015.

[120] Dominic J Packer. On being both with us and against us: A normative conflict model of dissent in social groups. *Personality and Social Psychology Review*, 12(1):50–72, 2008.

[121] M Afzalur Rahim. A measure of styles of handling interpersonal conflict. *The Academy of Management journal*, 26(2):368–376, 1983.

[122] M Afzalur Rahim. Toward a theory of managing organizational conflict. *International Journal of Conflict Management*, 13(3):206–235, 2002.

[123] M Afzalur Rahim. *Managing Conflict in Organizations*. Routledge, New Brunswick, 4th edition, 7 2017.

[124] Sarah T Roberts. Commercial content moderation: Digital laborers' dirty work. In Safiya Umoja Noble and Brendesha Tynes, editors, *The Intersectional Internet: Race, Sex, Class and Culture Online*, chapter Commercial, pages 147–160. NY: Peter Lang, New York, 2016.

[125] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, Shammur A Chowdhury, and Bernard J Jansen. A Literature Review of Quantitative Persona Creation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[126] Katrin Scheibe, Kaja J. Fietkiewicz, and Wolfgang G. Stock. Information behavior on social live streaming services. *Journal of Information Science Theory and Practice*, 4(2):6–20, 2016.

[127] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. Drawing from justice theories to support targets of online harassment. *New Media & Society*, 23(5):1278–1300, 2021.

[128] Sarita Schoenebeck, Carol F Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair after Online Harassment. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–18, 2021.

[129] Joseph Seering. Reconsidering Community Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 10 2020.

[130] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. Metaphors in moderation. *New Media & Society*, 24(3):621–640, 3 2022.

[131] Joseph Seering, Robert Kraut, and Laura Dabbish. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 111–125, 2017.

[132] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. Beyond Dyadic Interactions: Considering Chatbots as Community Members. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[133] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 7 2019.

[134] Kay Kyeongju Seo. Utilizing peer moderating in online discussions: Addressing the controversy between teacher moderation and nonmoderation. *American Journal of Distance Education*, 21(1):21–36, 5 2007.

[135] Tony L Simons and Randall S Peterson. Task conflict and relationship conflict in top management teams: The pivotal role of intragroup trust. *Journal of Applied Psychology*, 85(1):102–111, 2000.

[136] Leif Singer, Fernando Figueira Filho, Brendan Cleary, Christoph Treude, Margaret-Anne Storey, and Kurt Schneider. Mutual Assessment in the Social Programmer Ecosystem: An Empirical Investigation of Developer Profile Aggregators. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 103–116, 2013.

[137] Lee Sproull and Sara Kiesler. Reducing Social Context Cues: Electronic Mail in Organizational Communication. *Management Science*, 32(11):1492–1512, 1986.

[138] Tim Squirrell. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society*, 21(9):1910–1927, 9 2019.

[139] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21, 2019.

[140] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–14, 5 2021.

[141] Daniel J. Steinbock. Data Matching, Data Mining, and Due Process. *Georgia Law Review*, 40(1):1–86, 2005.

[142] Lucy Suchman. Making Work Visible. *Communication of the ACM*, 38(9):56–64, 1995.

[143] Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data*, 5(1):44, 2010.

[144] Thomas Theodoridis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Assessing the Reliability of Facebook User Profiling. In *Proceedings of the 24th International Conference on World Wide Web*, pages 129–130, 5 2015.

[145] Kenneth W. Thomas and Ralph H. Kilmann. Comparison of Four Instruments Measuring Conflict Behavior. *Psychological Reports*, 42(3):1139–1145, 1978.

[146] Stella Ting-Toomey. Conflict Face-Negotiation Theory. In *Conflict Management and Intercultural Communication*, pages 123–143. Routledge, 2 2017.

[147] Dean Tjosvold, Alfred S. H. Wong, and Paulina M. K. Wan. Conflict Management for Justice, Innovation, and Strategic Advantage in Organizational Relationships. *Journal of Applied Social Psychology*, 40(3):636–665, 2010.

[148] Catalina L. Toma, Jeffrey T Hancock, and Nicole B Ellison. Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, 34(8):1023–1036, 2008.

[149] Twitch.tv. Transparency Report 2020. Technical report, Twitch Interactive, Inc., CA, San Francisco, USA, 2020.

[150] Emily van der Nagel and Jordan Frith. Anonymity, pseudonymity, and the agency of online identity: Examining the social practices of r/Gonewild. *First Monday*, 20(3), 2 2015.

[151] Ruben van Wendel de Joode. Managing Conflicts in Open Source Communities. *Electronic Markets*, 14(2):104–113, 2004.

[152] Andreas Veglis. Moderation techniques for social media content. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8531, pages 137–148, 2014.

[153] Ting-Yu Wang, F Maxwell Harper, and Brent Hecht. Designing Better Location Fields in User Profiles. In *Proceedings of the 18th International Conference on Supporting Group Work*, pages 73–80, 11 2014.

[154] Deborah Weider-Hatfield and John D. Hatfield. Relationships Among Conflict Management Styles, Levels of Conflict, and Reactions to Work. *The Journal of Social Psychology*, 135(6):687–698, 1995.

[155] Michael Wenzel, Tyler G. Okimoto, Norman T. Feather, and Michael J. Platow. Retributive and restorative justice. *Law and Human Behavior*, 32(5):375–389, 2008.

[156] Monica T. Whitty. Liberating or debilitating? An examination of romantic relationships, sexual relationships and friendships on the Net. *Computers in Human Behavior*, 24(5):1837–1850, 2008.

[157] Donghee Y. Wohn. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[158] Donghee Y. Wohn and Guo Freeman. Audience Management Practices of Live Streamers on Twitch. In *Proceedings of the ACM International Conference on Interactive Media Experiences*, pages 106–116, 2020.

[159] Donghee Y. Wohn, Guo Freeman, and Caitlin McLaughlin. Explaining Viewers' Emotional, Instrumental, and Financial Support Provision for Live Streamers. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[160] Yu Chu Yeh. Analyzing online behaviors, roles, and learning communities via online discussions. *Educational Technology and Society*, 13(1):140–151, 2010.

[161] Amy X. Zhang and Justin Cranshaw. Making Sense of Group Chat through Collaborative Tagging and Summarization. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–27, 11 2018.

[162] Leizhong Zhang, Qiong Yang, Ta Bao, Dave Vronay, and Xiaoou Tang. Imlooking: Image-based Face Retrieval in Online Dating Profile Search. In *Extended Abstracts on Human Factors in Computing Systems*, pages 1577–1582, 2006.

[163] Chen Zhao and Gonglue Jiang. Cultural differences on visual self-presentation through social networking site profile images. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1129–1132, 2011.