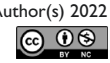
Article

new media & society
1–22



DOI: 10.1177/14614448221109804
journals.sagepub.com/home/nms


(In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit

Hibby Thach
University of Illinois Chicago, USA

Samuel Mayworm
Daniel Delmonaco
Oliver Haimson
University of Michigan, USA

Abstract

Research suggests that marginalized social media users face disproportionate content moderation and removal. However, when content is removed or accounts suspended, the processes governing content moderation are largely invisible, making assessing content moderation bias difficult. To study this bias, we conducted a digital ethnography of marginalized users on Reddit's /r/FTM subreddit and Twitch's "Just Chatting" and "Pools, Hot Tubs, and Beaches" categories, observing content moderation visibility in real time. We found that on Reddit, a text-based platform, platform tools make content moderation practices invisible to users, but moderators make their practices visible through communication with users. Yet on Twitch, a live chat and streaming platform, content moderation practices are visible in channel live chats, "unban appeal" streams, and "back from my ban" streams. Our ethnography shows how content moderation visibility differs in important ways between social media platforms, harming those who must see offensive content, and at other times, allowing for increased platform accountability.

Keywords

Content moderation, digital ethnography, marginalization, Reddit, social media platforms, Twitch, visibility

Corresponding author:
Hibby Thach, University of Illinois Chicago, Chicago, IL 60607, USA.
Email: kthach3@uic.edu

Introduction

In January 2022, Twitter confirmed it had permanently suspended US Representative Marjorie Taylor Greene's (R-Georgia) personal account for violating COVID-19 misinformation guidelines (Sullivan and Dale, 2022). This was not the first time Twitter had revealed its content moderation decisions; in January 2021, the platform suspended US President Donald Trump's personal account for policy violations (Twitter Inc, 2021). Yet despite these well-publicized peeks behind the curtain of social media content moderation, most content moderation remains hidden—often by design—from public view (Roberts, 2019). The implications of this are profound; although there is an undeniable public interest justification for private processing of harmful, violent, and illegal content, there is a real danger of moderators—whether individuals or institutions—transferring their own cultural biases to the content moderation process (Binns et al., 2017). When these decisions are biased, they other and marginalize certain groups of people (Haimson et al., 2021b). Mitigating this bias and marginalization may require increased visibility to increase possible accountability, but the extent to content moderation visibility's impact may differ across platforms.

Twitter, where Greene and Trump's suspensions occurred, functions as a top-down platform, meaning that most of its moderation occurs through platform moderators and/or algorithmic tools. On bottom-up platforms like Twitch and Reddit, the subjects of our study, most moderation occurs through volunteer moderators (Bradford et al., 2019). Twitch is a live streaming and live chat platform centered around game live streaming and participatory online communities (Hamilton et al., 2014). Reddit is a text-based platform where users can create or join topic-based discussion communities, including support communities for specific marginalized communities. While top-down platforms do moderate marginalized users' content disproportionately (Haimson et al., 2021b), bottom-up platforms offer a space to analyze content moderation practices outside of and alongside algorithmic tools. Yet sometimes, such as in the cases of Reddit and Twitch, a platform blurs the lines between top-down and bottom-up moderation by combining company-level guidelines with volunteer moderation, employing what Caplan (2018) called community-reliant approaches.

In this article, we examine content moderation in real-time across Reddit and Twitch to observe the differences in community-reliant approaches for content moderation visibility. We explore how content moderation visibility impacts marginalized users on these bottom-up platforms. We ask:

RQ1. How do social media bans of marginalized users and their content make visible places where content moderation systems and policies do not work as intended?

RQ2. How might digital ethnographic methods identify and track seemingly invisible content moderation practices?

To answer these research questions, we conducted a digital ethnography of Reddit's /r/FTM subreddit and Twitch's "Just Chatting" and "Pools, Hot Tubs, and Beaches" categories, two field sites that consist of particularly marginalized users, including

transgender users on Reddit (Alford, 2021) and female streamers (Ferrari, 2021) and streamers of color (Grayson, 2021b) on Twitch. We found that although Twitch and Reddit are both bottom-up platforms utilizing community-reliant approaches, their differences in platform features affected the content moderation visibility of each. On Reddit, platform tools make content moderation practices invisible to users, but moderators may make these practices visible through communication with users. An example is subreddit moderators' leaving explanatory "removal reason" comments on removed content, either manually writing these explanations or using AutoModerator (Reddit's native automated moderation tool) to do so automatically (Reddit, 2022a). On Twitch, content removal can be seen in channel live chats, "unban appeal" streams, and "back from my ban" streams.[1] Both platforms' content moderation practices are also made visible through press coverage of noteworthy bans and content removals, usually of public figures like celebrities and politicians.

A digital ethnography of Reddit and Twitch provides a glimpse of how marginalized populations experience content moderation on bottom-up platforms. Methodologically, our digital ethnography contributes to the existing literature by considering content moderation visibility holistically, combining analysis of content removal, account bans, and press coverage. Theoretically, this adds to previous literature on "following the field" (Burrell, 2009) by considering how ethnographic fieldwork interacts with larger phenomena like media attention. We hope that the insights from this study and their implications can help platforms improve content moderation practices to better serve marginalized populations. This article will not "solve" content moderation but make actionable suggestions toward possible solutions. Our study provides insight into new, unique ways of making content moderation visible while also discussing visibility's pitfalls alongside its benefits.

Related work

A brief overview of content moderation research and marginalized groups

Content moderation is a necessary part of the work platforms must do, as it enables platforms to protect users and user groups from one another and antagonists, allowing platforms to remove offensive and/or illegal content (Gillespie, 2017, 2018) However, discrimination and disproportionate moderation faced by marginalized groups are unanticipated consequences of platform moderation. Many journalists and scholars have argued that moderation practices across different platforms disproportionately target marginalized groups such as queer people and women of color, particularly queer/transgender Black women (Electronic Frontier Foundation, 2019; Salty, 2019; Smith et al., 2021), individuals with mental illness (Feuston et al., 2020), and many others (Haimson et al., 2021b; Lux and Lil Miss Hot Mess, 2017). Haimson et al. (2021b) reiterated the problems with false positives, where content and accounts are removed despite apparent adherence to site policies and community norms.

Making visible the practices that generate such potentially biased moderation is difficult (Feuston et al., 2020; Nakamura, 2015), as platforms tend to lack both transparency and accountability in their appeals processes (Vaccaro et al., 2020; West, 2018). Attempts

to generalize content moderation practices across all groups collapse contexts that can often negatively impact marginalized users (Caplan, 2018). For example, despite Facebook's claim to defend all races and genders equally (Angwin et al., 2017), marginalized users who rearticulate the hate speech directed at them in their effort to expose and confront the injustice face the same content removal as their antagonists (Jan and Dwoskin, 2017). The wide margins of error and the lack of context and nuance of current computational solutions to content moderation make such interventions prone to overcorrection and are thus inadequate for many users, especially for the most marginalized (see also Buolamwini and Gebru, 2018; Oliva et al., 2021).

Yet human moderation has its limitations, such as the potential psychological trauma of such work (Roberts, 2019). In addition, the algorithms designed to mitigate the human factor are themselves trained by humans who may hold biases that can be transferred to the moderation algorithms (Binns et al., 2017).

Visibility and content moderation

All steps of content moderation (flagging, deliberation, etc.) except the final steps of content removal and account suspension are invisible to most social media users (Gillespie, 2018). Roberts (2016) argued that content moderation and the humans who do it are *meant* to be invisible; by keeping the mechanisms of content moderation out of the public eye, platforms seek to project an image of objectivity (Roberts, 2019). In many cases, the extent of content moderation visibility is akin to Facebook's flagging and support dashboard, which still manages to keep the algorithmic parts of the process invisible (Crawford and Gillespie, 2014). On Tumblr, the 2018 NSFW (not suitable for work) content ban made its content moderation processes visible as well, allowing users and researchers to critique Tumblr's policies, processes, and politics (Sybert, 2021). On Twitch, acts of moderation can be both visible and invisible, as users can see moderation happening in real-time in Twitch's live chats but cannot see the behind-the-scenes coordination and sanctions enacted by channel moderators (Cai et al., 2021). In comparison, moderation on Reddit is a mostly invisible process; because Reddit is not a real-time platform, Reddit users cannot see content or account removals taking place as they happen (Jhaver et al., 2019a). As evidenced here, content moderation is difficult to research, as researchers often can analyze only what is seen (Gerrard and Thornham, 2020). Often, the invisible parts of content moderation can be researched only when those who experience content moderation or those enacting the moderation publicly describe their experiences.

Content moderation is made invisible through various actions. Shadowbanning, or the act of making users' content invisible to other users without removing it entirely (West, 2018), is one of these invisible processes that make it difficult to hold platforms responsible for erasing marginalized voices (Blunt and Wolf, 2020). Platforms usually say that they do not shadowban users (Gadde and Beykpour, 2018; Rosen, 2019; TeamYouTube, 2020), but multiple studies have shown the narratives of users who report experiencing some form of shadowbanning (Are, 2020; Middlebrook, 2020; Salty, 2019; Smith et al., 2021). Reporters (Biddle et al., 2020) found that TikTok leadership instructed moderators to disproportionately moderate posts created by "ugly, poor, or disabled" users. Such

discoveries make visible areas where content moderation is most hidden, allowing for critique and calls for clearer, more consistent, and more judicious content moderation practices. The complaints of marginalized users targeted by such moderation may be discounted, however, and the harassed users themselves might face retaliatory moderation as punishment, as when Black users talk explicitly about racism on Facebook (Guynn, 2019). In addition, appeals processes to dispute problematic moderation often have systematic failures that disenfranchise these marginalized groups from the process entirely (Vaccaro et al., 2020).

Reddit, Twitch, and content moderation

On Reddit, subreddit content moderation is predominantly performed by teams of volunteer moderators. Moderators focus on enforcing both subreddit-specific community guidelines and Reddit's sitewide guidelines that apply to all subreddits. Common Reddit moderator actions can include removing posts, comments, and media that violate subreddit guidelines and banning users who violate subreddit guidelines (Jhaver et al., 2019b; Reddit, 2022b). However, on Twitch, live chat means that content moderation practices happen much quicker. Cai and Wohn (2019) found that despite being a predominantly bottom-up platform, Twitch presents unique challenges to content moderation compared to asynchronous communities such as Wikipedia and Reddit. As in other bottom-up platforms, moderators for Twitch micro-communities (channels built around a single or group of streamers), who are charged with negotiating complex interpersonal relationships with streamers and viewers, are unpaid volunteers, without relevant purposeful training in their important moderator obligations (Cai and Wohn, 2019).[2]

Many volunteer moderators on both Reddit and Twitch do the work in place of algorithms because of their strong commitment to their communities (Seering et al., 2019). Seering et al. (2019) warned, however, that imbalance between platform-driven and user-driven governance can create safe spaces for extremist communities and hate groups to develop and thrive. However, Seering et al. (2017) also found that despite the potential for online communities to become spaces for hate, moderators can shape pro- and antisocial behaviors through their moderation practices and example setting. This is especially pertinent in the context of this study, as researchers argue Reddit contributes to "toxic technocultures," or cultures enabled and propagated through sociotechnical networks and online gaming centered around othering those perceived as outside its culture (Massanari, 2017). These spaces generally "demonstrate retrograde ideas of gender, sexual identity, sexuality, and race and push against issues of diversity, multiculturalism, and progressivism." (Massanari, 2017: 333) While a platform itself may not be "toxic," it can still allow substantial harassment and abuse toward marginalized people. Reddit and Twitch enable spaces for marginalized groups, but these spaces ultimately face increased harassment and abuse as their progressive and multicultural natures threaten the platforms' dominant culture.

As several examples, volunteer moderators on Reddit often witness troubling trends of racist, anti-Semitic, and misogynistic content posted on their subreddits, demonstrating the "default masculine whiteness" of Reddit's majority userbase (Gilbert, 2020). These observations mirror Reddit's history of hosting large, explicitly bigoted

user communities such as r/TheRedPill, /r/Incels, and other subreddits characterized by an overt hatred of people holding certain marginalized identities and a general rejection of "diversity, multiculturalism, and progressivism" (Gilbert, 2020; Massanari, 2017). Massanari (2017) cited Reddit's structural design, including elements such as the "upvote"-based content-visibility algorithm, user pseudonymity, and the ease of creating new user accounts, as key factors in the creation and success of these hate-based communities on Reddit. Reddit's sitewide administrators have banned some explicitly hateful or abusive subreddits in the past (Chandrasekharan et al., 2017). However, Habib et al. (2019: 3) found that these subreddit removals rarely take place "except in reaction to media pressure or catastrophic external events taking place as a result of discourse on [Reddit]," an approach that is criticized by some Reddit users for only taking place "after a significant amount of damage has already been observed."

We extend the "toxic technoculture" designation to communities on Twitch, the largest platform for streaming online gaming. While Twitch's guidelines present themselves as straightforward, "common sense" documents, they replicate and reinforce oppressive and discriminatory attitudes toward marginalized individuals (Gray and Leonard, 2018). Twitch's definition of sexual content is vague, subjective, and often contradictory, enabling further marginalization of streamers who are not cisgender heterosexual men (Ruberg, 2021; Zolides, 2020). Users are currently asking Twitch to improve their community guidelines and protections for marginalized streamers and viewers. For example, in August 2021, marginalized Twitch streamers rallied under the banner of #TwitchDoBetter, a Twitter hashtag created by streamer RekItRaven after a hate raid took over their community chat; for many of these streamers and creators, the hashtag expresses their frustration over Twitch's inadequate or misguided handling of harassment and abuse on their platform (Grayson, 2021b). In addition, Twitch has faced controversy related to sexual content on the platform, from yoga ASMR (autonomous sensory meridian response) streams (Diaz, 2021) to the rise of hot-tub streams (Asarch, 2021; Grayson, 2021a, 2021c).

Method

To examine content moderation on two different bottom-up platforms using community-reliant approaches—Reddit and Twitch—we conducted a 5-week digital ethnography (Hine, 2000) observing content moderation in real-time. During this digital ethnography, we took field notes on interactions between moderators and users on Reddit and between streamers, moderators, and viewers on Twitch for around 10 hours per week. We conducted our digital ethnography at a specific field site on each platform: the /r/FTM subreddit and the "Just Chatting" and "Pools, Hot Tubs, and Beaches" categories on Twitch. Our goals were to observe how content moderation of marginalized users and their content might illuminate faults in policies and systems (RQ1), and how digital ethnographic methods might better identify and make visible seemingly invisible content moderation practices (RQ2). All aspects of this study were reviewed and deemed exempt from oversight by University of Michigan's Institutional Review Board (IRB).

Digital ethnography

This study utilizes digital ethnographic methods to observe content moderation in real-time, analyzing the different features of a live chat and live streaming-based platform (Twitch) alongside the features of a solely text-based platform (Reddit). Digital ethnography refers to ethnography mediated by digital technologies, which can include ethnographic accounts of both offline and online groups (Murthy, 2008, 2011) For Reddit, we originally chose a variety of both discussion and humor-based subreddits that cater to transgender user communities, including /r/traaaaaaannnnnnnnns and /r/FTM. We later decided to focus specifically on /r/FTM because it is a large discussion-oriented community that presented more opportunities to witness visible moderation. For Twitch, we began with the "Pools, Hot Tubs, and Beaches" category as it had garnered press and controversy around bans and harassment of female streamers (Grayson, 2021a, 2021c) From there, we discovered a variety of streamers partaking in "unban appeal" videos, leading us to expand the field site to include any streamer streaming these types of videos. To find relevant videos, the authors examined recent streams in the "Pools, Hot Tubs, and Beaches" category and the "Just Chatting" category, as the latter became a category where "unban appeal" videos were frequently uploaded. As Feuston et al. (2020) similarly disclosed, our ethnography is naturally political but remains engaged with ethical responsibilities and considerations, as we use digital ethnography to connect and empathize with marginalized experiences within our field sites and overall to listen actively and adaptively (Winter and Lavis, 2020).

Data collection and analysis

We used Braun and Clarke's (2006) thematic analysis methods to qualitatively analyze our field-note data alongside excerpts from our field sites. Our team followed Braun and Clarke's (2006) six-step process of familiarizing ourselves with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and then producing our final report. Starting this six-step process, the first and second authors took field notes at least 3 days per week, reading and rereading our data while taking notes on initial ideas for open codes. We inductively coded lines and important text with our open codes and wrote memos following each day of coding. Following our initial coding, we met collaboratively as a team to separate and combine codes based on similarity and to identify emergent themes (Braun and Clarke, 2006). In further meetings, we followed this same process, reviewing and refining our themes on two levels (Braun and Clarke, 2006): The first involved making sure that our coded field-notes excerpts matched the themes we had come up with (and if not, refining our themes as needed to make sense with the data), and the second involved refining our final themes as needed to fit all relevant data. Then, we defined and named these final themes to reflect and represent the data. These themes included "actions that make moderation visible," "visibility outside the platform," "users' responses to moderation," and "invisible moderation." All these themes revolve around how content moderation becomes visible.

Moderation visibility exists on a continuum. While we were able to determine that moderation and removal had occurred, in most instances, the removed or altered content

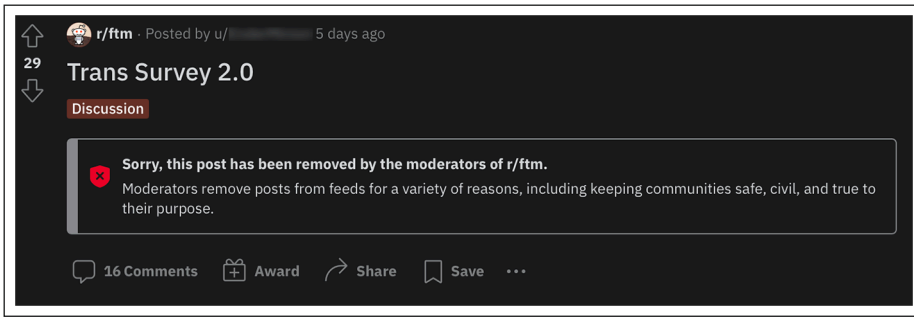


Figure 1. A Reddit post removed by the moderators of r/ftm.

itself was often rendered inaccessible. For instance, on Twitch, when one sees a comment removed by a moderator, that is a visible moment of content moderation. However, if one does not see the original comment and only sees that it was removed or moderated, there is a level of invisibility to this content moderation, as users do not know what the original message included. In addition, rewatching an uploaded stream will automatically hide any moderated comments so that viewers do not see comments moderated during the stream. Complexities that arise around content moderation's temporality make Twitch's live-chat and live-streaming features an interesting place to observe content moderation visibility.

"Real-time" moderation differs on asynchronous platforms like Reddit compared to synchronous platforms like Twitch, where content removals can be viewed as they take place. In comparison, "real-time" moderation on Reddit becomes visible when a user witnesses a Reddit post or comment replaced with a "removed by moderator" label while attempting to access or refresh the content (see Figure 1). Although this witnessing can take place almost immediately after a moderator removes content, it still introduces a level of invisibility to Reddit moderation, as Reddit users unaware of content before its removal may never realize content moderation has taken place. Although it is technically possible to find links to removed Reddit content, the process still relies on being aware content was removed. One way to access removed Reddit content is to view a moderator's public comment history and to access their comments on a post that was removed. While the body text of removed posts is invisible, it is possible to tell what kind of content was removed by reading the interactions between the moderator who removed the content and the user who originally posted it (see Figure 2). If there are no such interactions on that content, however, it is difficult to know what was removed.

Ethical considerations

In this article's write-up, we made decisions on how to report data using ethical fabrication (Markham, 2012) and anonymization because we were not able to get informed consent to participate in research from Twitch and Reddit users (franzke et al., 2020). For Reddit, we anonymized account usernames to maintain posters' privacy, as they are

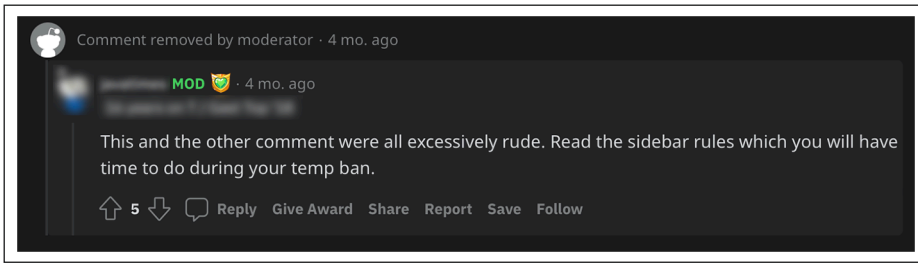


Figure 2. A Reddit moderator explaining why they removed a comment.

not public figures and often use Reddit to discreetly find information on trans-related issues. Unlike Twitch, Reddit posts and comments are typically visible and discoverable on search engines, along with the accounts that post them. Because of this risk, we modified Reddit users' post and comment text (e.g. paraphrasing) to reduce traceability back to their Reddit accounts. For Twitch, we did not anonymize streamers' usernames or quotes, as we consider them public figures when they broadcast content. However, when streamers reference viewers in the chat, we anonymized those people's usernames to protect their identities; no anonymization of viewers' quotes is needed, as moderated content is not viewable in an uploaded stream after the stream has been completed and is thus not traceable.

Limitations

Like any ethnography-based research, this work involves limitations. First, our data collection period (summer 2021) greatly influenced the types of content we saw on Twitch and Reddit. It was a time of mass hate raids on Twitch (Grayson, 2021a), the introduction of the "Pools, Hot Tubs, and Beaches" category, and mass transphobic hate raids on Reddit (Alford, 2021). These events made these platforms opportune spaces of inquiry for our study but may not reflect average amounts or intensities of problematic content. Second, our research was conducted in a mostly US context, as the research team is based solely in the United States. This means that our results may not be as applicable to non-US and non-Western contexts; however, our results provide important insights about these platforms nonetheless.

Results: (in)visible moderation

The relative visibility of content moderation processes differed between the two platforms studied. On Reddit, content moderation becomes most visible during direct, text-based comment interactions between the users and the moderators, which generally occur after a moderator removes a user's post or comments, and when Reddit users challenge the actions of subreddit moderators, through appeals or requests for explanation, when users' content is removed. On Twitch, streamers and viewers actively participate in the content moderation process or discuss content moderation decisions during live

streams, revealing their interventions with the phrase "message deleted by a moderator" and through "unban appeals," "mod apps" (communal review of applications to join the moderation team), and "unban" public announcement videos.

Content moderation visibility

The examples described in this section highlight important differences between platforms related to marginality and content moderation. On Reddit, we describe a case of content moderation revolving around efforts to protect marginalized groups. However, on Twitch, we describe a case where a marginalized streamer argues that she has been discriminated against by the platform's moderation. Content moderation visibility is integral in both cases. Reddit's mostly invisible content moderation processes help protect /r/FTM users from viewing abusive content. Yet invisible content moderation does not help marginalized Twitch streamers in the same way; instead, they are forced to make moderation visible as one way to right a platform's wrongs.

Actions that make moderation visible on Reddit. On Reddit, most moderator actions are unpublicized, unless the subreddit's moderation team specifically seeks input from their users about their decisions. Occasionally, users will post screenshots of their content that was removed by the moderators with a caption asking for an explanation or expressing frustration with the removal. While this kind of user action can make moderation visible, these kinds of posts are typically quickly removed by the moderators and can result in a user's being banned, making viewing the moderation actions in question quite difficult.

In early July, a wave of bot accounts with transphobic usernames and profile pictures began mass-following and sending abusive, often violent messages to openly trans users throughout Reddit. On /r/FTM, users began posting about being followed by these transphobic accounts. Users occasionally received transphobic direct messages from these accounts as well. While Reddit has a system to block other user accounts, this feature functions only to prevent the harassed user from seeing the blocked user's comments and posts. The blocked user can still see and interact with the harassed user's posts and comments and can also make another Reddit account to continue their harassment. Trans Reddit users expressed frustration, fear, and a sense of helplessness on the platform while this problem took place, particularly in response to the sheer lack of options users had to deal with these bots. There was little that moderators of trans-specific subreddits like /r/FTM could do about this problem; even if a moderator were to ban an abusive account from posting or commenting on a subreddit, that user could still view the subreddit and direct message its users. Several other trans-specific subreddits, such as r/MtF and r/transpassing, set their subreddits to "private" to combat this problem, meaning users who were not already subscribed to those subreddits and wanted access were required to message a moderator to request access. /r/FTM moderators chose not to set their subreddit to "private" and instructed users to avoid engaging with the transphobic accounts or posting any screenshots of the accounts to the subreddit instead. The moderators asked users to report these abusive accounts directly to the moderator team so they could be banned as quickly as possible and to limit any discussion of the problem to

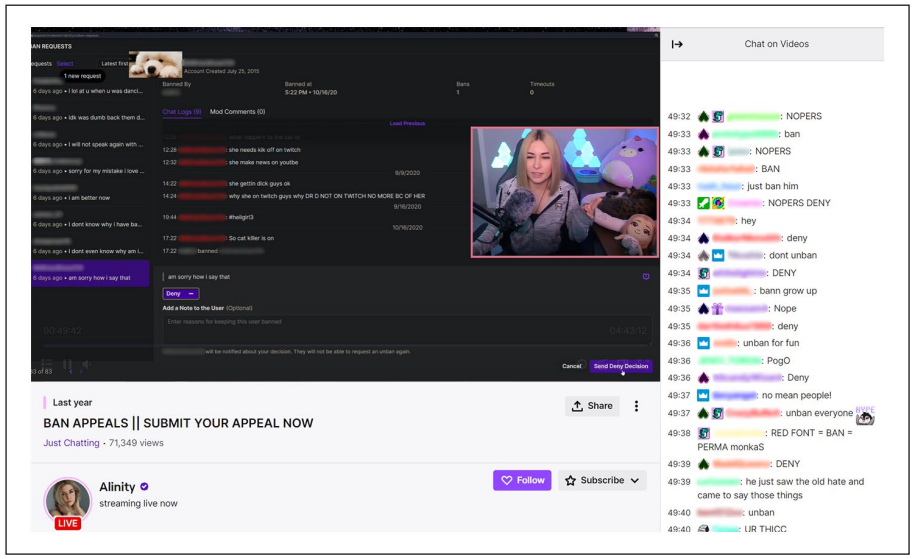


Figure 3. Twitch streamer Alinity doing an unban appeal stream and asking the chat if a user should remain banned or not.

a moderator-approved "megathread" about the topic. Reddit's site-wide administrative team initially did not remark on this wave of transphobic abuse.

Actions that make moderation visible on twitch. On Twitch, content moderation is made visible through the interactions between streamers and their moderators, the platform, and viewers. Streamers issue public announcements about their own ban or unban, submit unban appeals and/or mod apps on their live stream, and set norms by guiding viewers to follow live-chat rules. Often, streamers will poll moderators and/or viewers about the ban/unban status of a viewer and/or their content (see Figure 3). Streamers may also provide insight about certain bans or solicit explanations from a moderator on their team about a banning determination. In addition, streamers occasionally allow viewers to participate in the mod app process. Content moderation is especially evident when the streamer or their moderation team removes comments that break community guidelines (see Figure 4) and reiterates the rules, either on stream or in the chat.

The 2021 Twitch ban of streamer Sukesha Ray, an Asian American woman who streams to the "Just Chatting" and "Hot Tubs" categories, and her July return to the platform (after a month of being banned) exemplifies many of the attributes of the Twitch content moderation process and its potential for streamer-initiated visibility. Under the all-capitalized banner "WE SURVIVED THE INDEFINITE BAN!!!!! WERE BACK!!!!!!!," which she streamed in the "Just Chatting" and "Hot Tubs" spaces, Ray elicited a flood of comments from viewers asking about her ban, along with a corresponding response from moderators that "Sukes" herself would explain it when the stream officially started (she had been banned for exposing the aureole of her breast in a streamed video). Without the publicity generated by her unban appeal video and the July come-back stream, most Twitch users

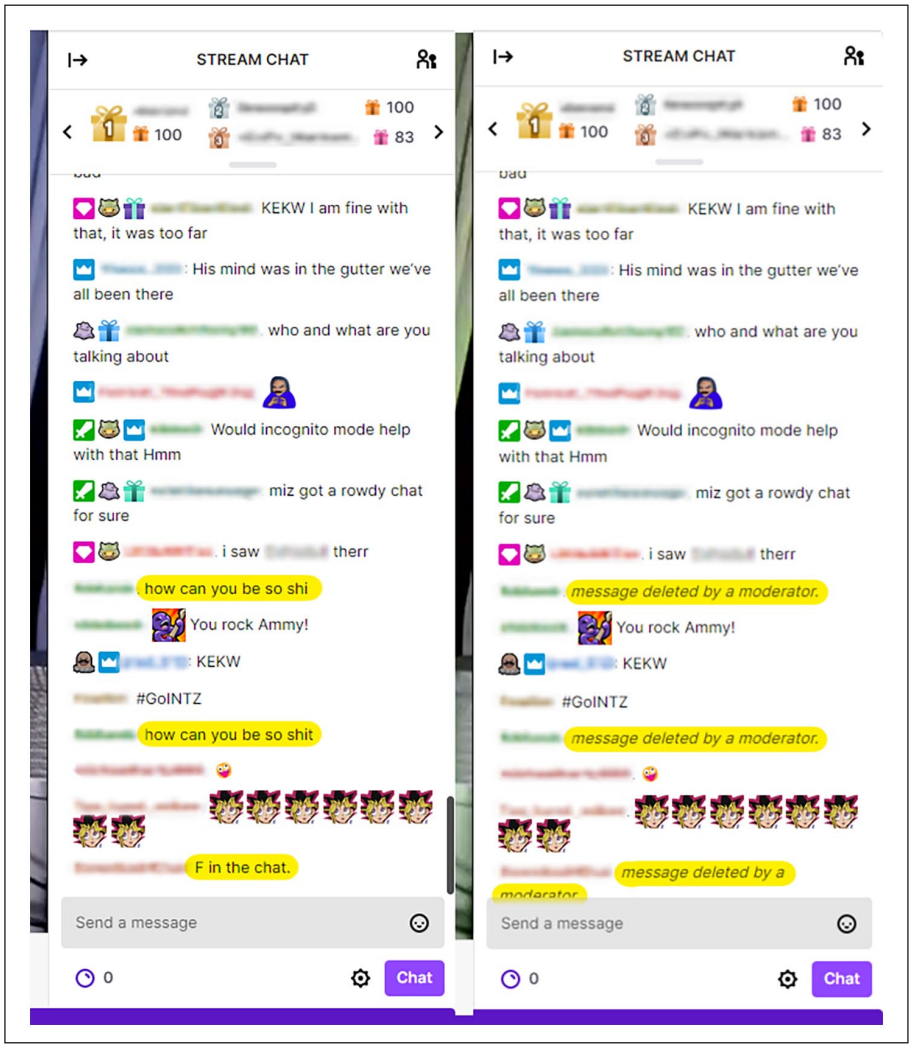


Figure 4. A side-by-side comparison of a Twitch channel's chat before and after comments were moderated.

would have been unaware of both the banning itself and the infraction that initiated it. Ray, speaking about how she took legal action and how she struggled to be officially reinstated, revealed to viewers both the inner workings of content moderation and the emotional toll that such actions can have on those who are moderated.

Responses to moderation

Users on Reddit. On /r/FTM, users commonly post about their frustration with moderation practices after experiencing content removal that they felt was either unjustified or

insufficiently explained. While /r/FTM users have the option of sending a direct message to the subreddit moderation team, many are either unaware of this option or deliberately choose to post their grievances on the subreddit to reveal moderator action. In some cases, when moderator contact is not forthcoming, other users, who have been made aware of a user's dissatisfaction, step in to explain the applicable rules and to support and guide the user to solutions or alternative approaches.

Comment removals on Reddit are generally more visible than post removals; comments removed by moderators have their text and usernames erased but do not entirely disappear like removed posts. Moderators sometimes go beyond simply reiterating the rule that had been broken, offering more extended responses in a comment section, taking the time, for example, to explain triggering language in a comment that the user had meant to be innocently humorous.

Viewers on Twitch. Twitch viewers' responses to content moderation are made visible through unban appeal videos. In our time spent watching Twitch streams in the "Just Chatting" and "Hot Tubs" categories, most viewer responses fell into two categories: denial and further harassment and genuine remorse. In the denial and further harassment category, viewers expressed disagreement with the content moderation decision and/or further violated the community guidelines, espousing more harassment and abuse in their unban appeal comment.

In most cases, bans were met with denial and/or further harassment in the ban appeals. Very few examples included "remorse," and we cannot determine if viewers who apologized were actually remorseful. However, what is important here is how these unban appeals make moderation decisions visible. Not only are the moderators' ban decisions revisited and reevaluated; they are done so publicly, and other viewers' opinions are taken into account. For example, one streamer doing an unban appeal stream asked viewers in the chat to post an angel or a devil emoji to decide whether a viewer should remain banned. The angel symbolizes forgiveness, suggesting that the viewer should be unbanned, while the devil symbolizes punishment, suggesting that the viewer should remain banned. This communal form of content moderation is a unique finding from this study, which future work should examine further.

Visibility outside of the platform

While individual instances of content moderation are important to analyze, they do not exist solely on the platforms they take place on. It is important to consider media attention paid to content moderation on Reddit and Twitch during our digital ethnography, as outside visibility contributes to total content moderation visibility. In addition, we must consider the timing of our digital ethnography alongside the simultaneous anti-trans, misogynist, and racist cultural climate. During our study, a mass of transphobic activity took place on Reddit, along with substantial misogynistic and racist activity on twitch. This section details this increased media attention alongside the contributions of media attention to content moderation visibility.

Press on Reddit. On July 9, the news platform *Jezebel* posted a detailed article about the wave of transphobic platform abuse on Reddit, featuring quotes from several trans Reddit

users (including moderators of trans-specific subreddits) about their experiences dealing with harassment. The article illustrated how the transphobic bot accounts exploited loopholes in Reddit's user ban and user blocking features and how little power trans Reddit users have in preventing or avoiding further abuse (Alford, 2021). Reddit administration took greater action against these transphobic accounts after the *Jezebel* article was published, introducing restrictions for new account usernames and promising to introduce further improvements to the user blocking/following features near the end of 2021. They also posted a user safety update encouraging affected users to turn off in-app/email notifications for new followers and to block accounts that post abusive content. In January 2022, Reddit announced an update to the user blocking feature that prevents blocked users from viewing activity from accounts that block them (Reddit, 2022c).

Press on Twitch. Twitch's "Pools, Hot Tubs, and Beaches" category has garnered wide press coverage as it officially became a category in May 2021 (Twitch, 2021). The category was introduced after a long history of vague guidelines around sexually suggestive content on Twitch (Grayson, 2021a, 2021c; Ruberg, 2021; Zolides, 2020). Since its induction, the category has included many popular streamers (such as Amouranth and IndieFoxx) who reported being banned from the platform and also described a lack of care from Twitch itself, claiming that the platform bans streamers for their content while remaining generally unresponsive to harassment and abuse that the streamers themselves face from viewers (Diaz, 2021). In other cases, marginalized, less-known streamers have reported difficulties appealing their bans, which they attribute to their lack of visibility compared to popular streamers. In addition, some Twitch streamers report that they experience targeted harassment, such as AnneAtomics, a white trans woman who claimed that the platform and/or its users may not want a trans woman to be a prominent hot tub streamer (Nightingale, 2021).

Discussion

We described social media content moderation's visibility by way of two digital ethnographies of online communities on two platforms with vastly different features. On Reddit, content moderation typically becomes visible when a subreddit's moderation team chooses to openly discuss their moderation decisions with their users through subreddit posts and comments. On Twitch, live video and live chat streams and streamers discussing or visibly moderating content in real-time make content moderation visible not just to moderators, but to users as well. In both cases, visibility is needed to increase accountability and to address concerns around unfair content removal or account suspension. However, also in both cases, visibility is less critical in instances where content obviously violates community guidelines, particularly abusive and harmful content.

Platform features and content moderation visibility

Reddit has few platform tools to make subreddit moderators' content moderation actions visible to its users. While subreddit moderators can view removed content and banned users via the moderation log, this information is invisible to Reddit users. As we described

in the "Results" section, moderation on /r/FTM typically became visible when a user posted an open objection to a moderation decision, which led to subreddit moderators entering the thread and explaining their decision in the comments. Moderation also became visible when the moderation team chose to openly discuss the subreddit guidelines with the user community, often asking for community input on major guideline changes. These unique user-moderator dynamics can take place on discussion-heavy support communities like /r/FTM with community moderators willing to openly discuss their decisions with their users. Previous literature has explored the frustration and perceptions of "unfairness" experienced by Reddit users whose content is removed without a moderator explanation or without notifying the user (Jhaver et al., 2019a). We expanded on these prior works by exploring the relationship between a marginalized user community on Reddit and their subreddit moderators, finding similarities in how moderation becomes visible in that community and how its users perceive moderation. We also explored how /r/FTM's moderators combatted Reddit's "toxic technoculture," (Massanari, 2015) including their limited power to combat transphobic abuse from outside the subreddit compared to Reddit's sitewide administrators (Alford, 2021; Habib et al., 2019). The actions taken by /r/FTM's moderators during the wave of sitewide transphobic abuse across Reddit were visible not only in and of itself but made the lack of moderation actions coming from the platform visible as well.

As seen in our Twitch ethnography of the "Just Chatting" and "Pools, Hot Tubs, and Beaches" categories, a platform with the capabilities to stream live content with live chat allows for increased content moderation visibility, both from the platform and from its streamers who moderate their live chats. It is important to consider how content moderation can be visible on all levels and which actors do the work to make this visible. Twitch, like most platforms, does not overtly make content moderation practices visible. Instead, the streamers described in our digital ethnography increase content moderation's visibility. We described how streamers spoke about Twitch as a higher authority and asked for the platform to do better, regardless of whether Twitch representatives were part of the audience hearing these appeals. However, the streamers provide details that viewers would otherwise never know about Twitch's content moderation practices and simultaneously interact with viewers while doing so. This differs from platforms like Reddit, where both subreddit moderation and users' discussion and criticism of subreddit moderation is less visible by design.

The future of (in)visible moderation

External visibility of platforms' content moderation practices is essential to building public awareness of marginalized populations' disproportionate moderation. It was from media external to the platforms themselves, for example, that drew attention to TikTok's suppression of disabled, fat, queer, and Black users (Biddle et al., 2020; Botella, 2019; Gebel, 2020; Köver, 2019) and to Instagram's content moderation practices disproportionately targeting women of color, queer people, and pole dancers (Are, 2020; Rodriguez, 2019; Salty, 2019; Smith et al., 2021). Increased transparency in content moderation practices is one common recommendation to improve users' online content moderation experiences (Haimson et al., 2021a; Jhaver et al., 2019b; Suzor

et al., 2019). Lack of transparency disproportionately impacts marginalized users, particularly when content falls into "gray areas" that cannot easily be determined as permissible or not according to guidelines (Haimson et al., 2021b). Making content moderation policies and decisions visible to users, especially marginalized users, might allow them to appeal to content moderation decisions more easily or understand why and how specific content was removed. In addition, including media attention in this study is central to our methodological contribution to digital ethnographic studies, as we must follow the fieldwork everywhere it goes, whether that be on the platform or in press about the platform.

To achieve more equitable social media content moderation for marginalized groups, we must consider the content moderation visibility that platforms' features allow and where and when content moderation should be visible and when it should not. For example, the abusive and harmful comments directed at marginalized Twitch streamers and Reddit users would do less harm if they had been less visible. Visibility on its own does not constitute change, but with visibility, powerful actors can be made aware of injustice and oppression. Moderation is an expression of power, which is continually renegotiated between the moderator and the moderated. With increased visibility, the moderated can appeal to the moderator and question structures of power. Platforms attempt to moderate abusive and harmful language and actions to maintain safety and neutrality for the everyday user, but as Roberts (2019) found, moderators undertake substantial labor to maintain these spaces. Although Twitch streamers call out inconsistencies and unfair decisions when Twitch removes their content or suspends their accounts, streamers still experience harm from seeing abusive and harmful content. Having to police their bodies even further, women, especially women of color, are disadvantaged by Twitch's community guidelines (Ruberg, 2021; Uttarapong et al., 2021; Zolides, 2020). Although a streamer like Sukesha Ray *did* accidentally have her areola slip out, other viewers in the comments and Sukesha Ray herself commented on how other streamers "do worse" but are more valuable to Twitch, so do not get their content removed. Instances of content moderation false positives or gray areas (Haimson et al., 2021b) like with Sukesha Ray need to be made visible so that marginalized people's disproportionate removals can be acknowledged and resolved.

We argue that an ideal world of content moderation allows marginalized users to call out inconsistencies and unfair treatment regarding content removal and account suspensions while also enabling automated and tiered governance systems that allow moderators and users to minimize their encounters with abusive and harmful content. This also requires that automated content moderation systems are improved so that they do not disproportionately target marginalized people. New features like Twitch's new chat filters and "channel-level ban-evasion detection" (detecting bots or accounts that try to bypass bans repeatedly; D'Anastasio, 2021) or Reddit's "opt out of followers" feature[3] (Reddit, 2021b) are constructive and positive developments, but only time will tell if their implementation meaningfully reduces harassment. Finding the right balance between protection and open expression is a difficult task. But the responsibility for developing and implementing processes, procedures, and technologies to maintain that delicate balance lies not with a platform's users but with the platforms themselves. Only by supporting adequate, judicious visibility of their content moderation processes can platforms evolve to meet this challenge.

Acknowledgements

We would like to thank Kim Greenwell's editing team for editing help on this paper, the UMSI graphics center for editing Figures 3 and 4 for us, and the members of the Social Transition Tech Lab at UMSI for their feedback on this work throughout the research and writing process.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science Foundation grant #1942125.

ORCID iDs

Hibby Thach https://orcid.org/0000-0002-0520-0218

Oliver Haimson https://orcid.org/0000-0001-6552-4540

Notes

1. "Unban appeal" streams consist of streamers discussing users banned from their channel's appeals, often asking for community opinions. "Back from my ban" streams consist of streamers sharing their return to Twitch after being unbanned, which may or may not include their recounting of the entire process.
2. As Hamilton et al. (2014) discussed, "user" is a confusing category on Twitch. We use "streamer" to refer to the individual live streaming on Twitch, whereas "viewer" refers to the individual viewing Twitch streams and participating in the live chat, and "moderator" refers to a streamer's moderation team.
3. On August 25, 2021, Reddit (2021a, 2021b) officially implemented the ability for Reddit users to prevent other users from "following" their accounts, hoping to improve user privacy and discourage harassment from abusive users.

References

Alford E (2021) How transgender Redditors are being driven from the site. *Jezebel*, 7 September. Available at: https://jezebel.com/transgender-redditors-are-being-driven-from-the-site-by-1847256024

Angwin J, ProPublica and Grassegger H (2017) Facebook's secret censorship rules protect white men from hate speech but not black children. *Propublica*, 28 June. Available at: https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

Are C (2020) How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist Media Studies* 20(5): 741–744.

Asarch S (2021) Hot tub livestreamers broadcasting in bikinis on Twitch have divided the platform's community. *Insider*. Available at: https://www.insider.com/twitch-hot-tub-meta-controversy-explained-2021-5

Biddle S, Ribeiro PV and Dias T (2020) Invisible censorship: TikTok told moderators to suppress posts by "ugly" people and the poor to attract new users. *The Intercept*, 16 March. Available at: https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/

Binns R, Veale M, Van Kleek M, et al. (2017) Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In: Ciampaglia GL, Mashhadi A and Yasseri T (eds) *Social Informatics*. Berlin: Springer, pp. 405–415.

Blunt D and Wolf A (2020) Erased: the impact of FOSTA-SESTA and the removal of backpage. *Hacking//hustling*. Available at: https://hackinghustling.org/erased-the-impact-of-fosta-sesta-2020/

Botella E (2019) TikTok admits it suppressed videos by disabled, queer, and fat creators. *Slate*, 4 December. Available at: https://slate.com/technology/2019/12/tiktok-disabled-users-videos-suppressed.html

Bradford B, Grisel F, Meares TL, et al. (2019) Report of the Facebook data transparency advisory group. The Justice Collaboratory, Yale Law School, New Haven, CT, April.

Braun V and Clarke V (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2): 77–101.

Buolamwini J and Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. *Proceeding of Machine Learning Research* 81: 77–91.

Burrell J (2009) The field site as a network: a strategy for locating ethnographic research. *Field Methods* 21(2): 181–199.

Cai J and Wohn DY (2019) Categorizing live streaming moderation tools: an analysis of Twitch. *International Journal of Interactive Communication Systems and Technologies* 9(2): 36–50.

Cai J, Wohn DY and Almoqbel MY (2021) Moderation visibility: mapping the strategies of volunteer moderators in live streaming micro communities. In: *Proceedings of IMX 2021: ACM international conference on interactive media experiences*, 2123 June, New York, pp. 61–72.

Caplan R (2018) *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches*. New York: Data & Society Research Institute.

Chandrasekharan E, Pavalanathan U, Srinivasan A, et al. (2017) 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: 31.

Crawford K and Gillespie T (2014) What is a flag for? Social media reporting tools and the vocabulary of complaint. *new media & society* 18(3): 410–428.

D'Anastasio C (2021) Twitch sues users over alleged "hate raids" against streamers. *Wired*, 10 September. Available at: https://www.wired.com/story/twitch-sues-users-over-alleged-hate-raids/

Diaz A (2021) Twitch bans Amouranth and Indiefoxx after yoga ASMR streams. *Polygon*, 21 June. Available at: https://www.polygon.com/22543478/twitch-amouranth-indiefoxx-yoga-asmr-stream-hot-tub-meta-ban

Electronic Frontier Foundation (2019) *EFF Project Shows How People are unfairly "Tossed Out" by Platforms' Absurd Enforcement of Content Rules*. San Francisco, CA: Electronic Frontier Foundation.

Ferrari P (2021) Twitch's #BikiniGate as seen by streamers: "It's just to criticize women.." *Madmoizelle*, 25 May. Available at: https://www.madmoizelle.com/le-bikinigate-vu-par-les-streameuses-cest-juste-pour-critiquer-les-femmes-1127836

Feuston JL, Taylor AS and Piper AM (2020) Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4: 40.

franzke as, Bechmann A, Zimmer M, et al. (2020) *Internet Research: Ethical Guidelines 3.0*. Chicago, IL: Association of Internet Research.

Gadde V and Beykpour K (2018) Setting the record straight on shadow banning. *Twitter Blog*, 26 July. Available at: https://blog.twitter.com/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning

Gebel M (2020) Black creators say TikTok still secretly hides their content. *Digital Trends*, 21 July. Available at: https://www.digitaltrends.com/social-media/black-creators-claim-tiktok-still-secretly-blocking-content/

Gerrard Y and Thornham H (2020) Content moderation: social media's sexist assemblages. *new media & society* 22(7): 1266–1286.

Gilbert SA (2020) "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a public scholarship site on Reddit: a case study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* 4: 19.

Gillespie T (2017) Governance of and by platforms. In: Burgess J, Poell T and Marwick A (eds) *SAGE Handbook of Social Media*. Thousand Oaks, CA: SAGE, pp. 254–278.

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

Gray KL and Leonard DJ (eds) (2018) *Woke Gaming: Digital Challenges to Oppression and Social Injustice*. Seattle, WA: University of Washington Press.

Grayson N (2021a) Hot tub streams are waning, but some people still think they're destroying Twitch. *Kotaku*, 23 April. Available at: https://kotaku.com/hot-tub-streams-are-waning-but-some-people-still-think-1846749371

Grayson N (2021b) Marginalized streamers beg Twitch to "do better" in wake of hate raids, poor pay. *The Washington Post*, 11 August. Available at: https://www.washingtonpost.com/video-games/2021/08/25/twitch-hate-raids-streamers-discord-cybersecurity/

Grayson N (2021c) Twitch's "Hot Tub Meta" has sparked off yet another debate about women's attire. *Kotaku*, 2 April. Available at: https://kotaku.com/twitchs-hot-tub-meta-has-sparked-off-yet-another-debate-1846600932

Guynn J (2019) Facebook while black: users call it getting "Zucked," say talking about racism is censored as hate speech. *USA Today*, 24 April. Available at: https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/

Habib H, Musa MB, Zaffar F, et al. (2019) To act or react: investigating proactive strategies for online community moderation. *arXiv*. Available at: https://arxiv.org/abs/1906.11932

Haimson OL, Dame-Griff A, Capello E, et al. (2021a) Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist Media Studies* 21(3): 345–361.

Haimson OL, Delmonaco D, Nie P, et al. (2021b) Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5: 466.

Hamilton WA, Garretson O and Kerne A (2014) Streaming on twitch: fostering participatory communities of play within live mixed media. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Toronto, ON, Canada, 26 April–1 May, pp. 1315–1324. New York: Association for Computing Machinery.

Hine C (2000) *Virtual Ethnography*. London; Thousand Oaks, CA: SAGE.

Jan T and Dwoskin E (2017) A white man called her kids the n-word. Facebook stopped her from sharing it. *The Washington Post*, 31 July. Available at: https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html

Jhaver S, Appling DS, Gilbert E, et al. (2019a) "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: 192.

Jhaver S, Bruckman A and Gilbert E (2019b) Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3: 150.

Köver C (2019) Discrimination: TikTok curbed reach for people with disabilities. *Netzpolitik*, 2 December. Available at: https://netzpolitik.org/2019/discrimination-tiktok-curbed-reach-for-people-with-disabilities/

Lux D and Lil Miss Hot Mess (2017) Facebook's hate speech policies censor marginalized users. *Wired*, 14 August. Available at: https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/

Markham A (2012) Fabrication as ethical practice. *Information, Communication & Society* 15(3): 334–353.

Massanari AL (2015) *Participatory Culture, Community, and Play: Learning from Reddit*. Bern: Peter Lang.

Massanari AL (2017) #Gamergate and the fappening: how Reddit's algorithm, governance, and culture support toxic technocultures. *new media & society* 19(3): 329–346.

Middlebrook C (2020) *The Grey Area: Instagram, Shadowbanning, and the Erasure of Marginalized Communities*. Rochester, NY: Social Science Research Network.

Murthy D (2008) Digital ethnography: an examination of the use of new technologies for social research. *Sociology* 42(5): 837–855.

Murthy D (2011) Emergent digital ethnographic methods for social research. In: Hesse-Biber SN (ed.) *Handbook of Emergent Technologies in Social Research*. Oxford: Oxford University Press, pp. 158–179.

Nakamura L (2015) The unwanted labour of social media: women of colour call out culture as venture community management. *New Formations* 86: 106–112.

Nightingale E (2021) Hot tub Twitch streamer banned after transphobic pile-on slams platform for failing creators. *PinkNews*, 13 July. Available at: https://www.pinknews.co.uk/2021/07/13/twitch-trans-streamer-anne-atomic-2/

Oliva TD, Antonialli DM and Gomes A (2021) Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture* 25(2): 700–732.

Reddit (2021a) R/changelog safety update on Reddit's follow feature. *Reddit*, 14 July. Available at: https://www.reddit.com/r/changelog/comments/ok8xzb/safety_update_on_reddits_follow_feature/ (accessed 22 April 2022).

Reddit (2021b) R/changelog you can now opt-out of being followed. *Reddit*, 25 August. Available at: https://www.reddit.com/r/changelog/comments/pbjshi/you_can_now_optout_of_being_followed/ (accessed 22 April 2022).

Reddit (2022a) AutoModerator. *Reddit*. Available at: https://mods.reddithelp.com/hc/en-us/articles/360002561632-AutoModerator (accessed 22 April 2022).

Reddit (2022b) General moderation info. *Reddit*. Available at: https://mods.reddithelp.com/hc/en-us/categories/360000255232-General-Moderation-Info (accessed 22 April 2022).

Reddit (2022c) R/blog – announcing blocking updates. *Reddit*, 18 January. Available at: https://www.reddit.com/r/blog/comments/s71g03/announcing_blocking_updates/ (accessed 22 April 2022).

Roberts ST (2016) Commercial content moderation: digital laborers' dirty work. In: Noble SU and Tynes BM (eds) *The Intersectional Internet: Race, Sex, Class and Culture Online*. Bern: Peter Lang, p. 12.

Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

Rodriguez J (2019) Instagram apologizes to pole dancers after hiding their posts. *CTVnews*, 6 August. Available at: https://www.ctvnews.ca/sci-tech/instagram-apologizes-to-pole-dancers-after-hiding-their-posts-1.4537820?cache=

Rosen G (2019) Remove, reduce, inform: new steps to manage problematic content. *Meta*, 10 April. Available at: https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/

Ruberg B (2021) "Obscene, pornographic, or otherwise objectionable": biased definitions of sexual content in video game live streaming. *new media & society* 23(6): 1681–1699.

Salty (2019) Exclusive: an investigation into algorithmic bias in content policing on Instagram. *Salty*, 23 October. Available at: https://twitter.com/saltyworldbabes/status/1313579751546728449

Seering J, Kraut R and Dabbish L (2017) Shaping pro and anti-social behavior on Twitch through moderation and example-setting. In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing (CSCW'17)*, Portland, OR, 25 February–1 March, pp. 111–125. New York: ACM.

Seering J, Wang T, Yoon J, et al. (2019) Moderator engagement and community development in the age of algorithms. *new media & society* 21(7): 1417–1443.

Smith SL, Haimson OL, Fitzsimmons C, et al. (2021) Censorship of marginalized communities on Instagram. *Salty*, 29 September. Available at: https://twitter.com/Saltyworldbabes/status/1443068587170803712

Sullivan D and Dale D (2022) One of Marjorie Taylor Greene's verified Twitter accounts permanently suspended from Twitter. *CNN*, 2 January. Available at: https://edition.cnn.com/2022/01/02/politics/marjorie-taylor-greene-twitter-suspension/index.html

Suzor NP, West SM, Quodling A, et al. (2019) What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication* 13: 18.

Sybert J (2021) The demise of #NSFW: contested platform governance and Tumblr's 2018 adult content ban. *new media & society*. Epub ahead of print 26 February. DOI: 10.1177/1461444821996715.

TeamYouTube (2020) @Herclueless we don't shadowban channels. . .https://t.co/f25cOgmwRV [Tweet]. *@Teamyoutube*, 22 October. Available at: https://twitter.com/TeamYouTube/status/1319372516398452737?ref_src=twsrc%5Etfw

Twitch (2021) Let's Talk About Hot Tub Streams. *Twitch Blog*. Available at: https://blog.twitch.tv/en/2021/05/21/lets-talk-about-hot-tub-streams

Twitter Inc (2021) Permanent suspension of @realDonaldTrump. *Twitter Blog*, 8 January. Available at: https://blog.twitter.com/en_us/topics/company/2020/suspension

Uttarapong J, Cai J and Wohn D (2021) Harassment experiences of women and LGBTQ live streamers and how they handled negativity. In: *ACM international conference on interactive media experiences*, 2123 June, pp. 7–19. New York: Association for Computing Machinery.

Vaccaro K, Sandvig C and Karahalios K (2020) "At the end of the day Facebook does what it wants": how users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4: 167.

West SM (2018) Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *new media & society* 20(11): 4366–4383.

Winter R and Lavis A (2020) Looking, but not listening? Theorizing the practice and ethics of online ethnography. *Journal of Empirical Research on Human Research Ethics* 15(1–2): 4366–4383.

Zolides A (2020) Gender moderation and moderating gender: sexual content policies in Twitch's community guidelines. *new media & society* 23(10): 2999–3015.

Author biographies

Hibby Thach is a master's student in the Department of Communication at the University of Illinois, and an incoming PhD student at the University of Michigan's School of

Information (UMSI). Their research explores identity, content moderation, and digital cultures within online communities and gaming spaces. Thach also researches trans technologies and representation within video games.

Samuel Mayworm is a graduate of the University of Michigan School of Information (UMSI), with a concentration on libraries and digital preservation. He researches marginalized users' experiences on social media and with content moderation, focusing on folk theorization among marginalized social media users.

Daniel Delmonaco is a PhD candidate at the University of Michigan School of Information (UMSI), where they are also affiliated with the Science, Technology, and Society Graduate Program and a Graduate Affiliate with the Center for Ethics, Society, and Computing. They research the online information seeking experiences of LGBTQ+ youth with a focus on sexual health and sex education. Delmonaco also researches marginalized people's experiences of content moderation on social media and healthcare providers' use of social media.

Oliver L Haimson is an Assistant Professor at University of Michigan School of Information (UMSI). He conducts social computing research focused on envisioning and designing trans technologies, and social media content moderation and marginalized populations, with a research goal of impacting technological inclusion of marginalized users. Haimson is a recipient of a National Science Foundation CAREER award.