



Article

Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms

new media & society
2018, Vol. 20(11) 4366–4383
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1461444818773059
journals.sagepub.com/home/nms



Sarah Myers West 

University of Southern California, USA

Abstract

Social media platforms play an increasingly important civic role as platforms for discourse, where we discuss, debate, and share information. This article explores how users make sense of the content moderation systems social media platforms use to curate this discourse. Through a survey of users ($n=519$) who have experienced content moderation, I explore users' folk theories of how content moderation systems work, how they shape the affective relationship between users and platforms, and the steps users take to assert their agency by seeking redress. I find significant impacts of content moderation that go far beyond the questions of freedom of expression that have thus far dominated the debate. Raising questions about what content moderation systems are designed to accomplish, I conclude by conceptualizing an educational, rather than punitive, model for content moderation systems.

Keywords

Accountability, content moderation, free expression, social media, survey, transparency, user studies

Social media platforms, including Facebook, Twitter, Instagram, and YouTube, among others, play an increasingly important civic role as platforms for discourse. They collectively create space for members of the public to gather, discuss, debate, and share information, a role that many platforms cultivate through their rhetoric: for example,

Corresponding author:

Sarah Myers West, Annenberg School for Communication and Journalism, University of Southern California, 1607 Blake Ave, Los Angeles, CA 90031, USA.
Email: sarahmye@usc.edu

Twitter's former CEO has referred to the platform as a "Global Town Square" (Baker, 2013). But these spaces are private in the sense of their commercial ownership; while platforms may serve as an important locus for individual self-expression and collective association, the users who populate them have relatively little influence on their architecture and governance. Social media platforms are

not simply about facilitating user-produced content across networks to large audiences or "end-users"; rather, they are primarily concerned with establishing the technocultural conditions within which users can produce content and within which content and users can be re-channelled through techno-commercial networks and channels. (Langlois et al., 2009)

Platforms may best be understood as sociotechnical assemblages and complex institutions, many of which are commercial but may be commercial in different ways (Gillespie, 2017). This article seeks to engage this dynamic more deeply by parsing the nuances of the systems platforms deploy to moderate the content through the perspective of their users.

Although platforms experience similar commercial and regulatory imperatives, social media companies may adopt differing strategies for the kinds of discursive communities they seek to cultivate. Several companies have adopted a rhetoric of free speech advocacy—for example, Twitter's UK General Manager Tony Wang once called the company "the free speech wing of the free speech party" (Halliday, 2012). Others frame themselves in the context of a global village, as Facebook did when it claimed to be a diverse "community of more than one billion people" (Facebook, 2016). These depictions suggest differing orientations around the company's obligation toward policing user content: a "free speech"-oriented company may be more likely to design its policies around defending free expression, while a "community"-oriented company may place greater focus on fostering good behavior among users and curbing harassment.

The term platform itself can serve as an obfuscating device, operationalized at times by social media companies in order to divert tensions between their obligations to multiple constituencies (Gillespie, 2010). For example, although a social media company may have an interest in free expression that enables users to post as much content as possible, it may not desire the kinds of expression that scare away advertisers. Or it may seek to balance the need to maintain the perception of being an open platform with demands by governments to police certain kinds of content. The discourse of platforms can enable them to perform an impartiality impossible to achieve in practice (Gillespie, 2017). It also presents the companies as unified entities, rather than as complex organizations with teams of employees that may have different cultures, values, and incentives with regard to public expression that collectively may or may not have coherence.

This tension—between community and marketplace, public sphere and private platform—is particularly apparent in companies' content moderation systems, which are designed to place bounded limits on undesirable forms of expression while maximally encouraging users to produce and post content. But platform governance processes are not one-way systems; users are implicated in systems designed to moderate content and are critical to their success or failure. This article adds a user-centered perspective to the growing body of literature on content moderation, considering how these strategies shape the expectations we form of companies' behavior by examining users' experiences

with the content moderation process. Through a survey of users ($n=519$) who have experienced a content removal, I explore how users interpret the role and function of companies in moderating their content, what kinds of content are taken down, and the impact this has on their public expression. I take as my starting point that users play a critical role both in the production of the content on offer and in the selection of which content should be removed, but at present we lack a clear conceptual framework for understanding content moderation as a relational process that incorporates both company systems and resources and user labor. The accounts I analyzed suggest that taking a user-focused approach to studying content moderation is a generative space for raising questions about what content moderation systems are themselves designed for.

The structure of this article proceeds as follows: first, I provide basic definitions for content moderation and explain how moderation systems are designed to work. Then, I describe what information is available to users seeking to understand a company's content guidelines, focusing on two types of documentation in particular, the company's terms of service and community guidelines. After outlining the methods for the study and overviews of its findings, I conclude by analyzing how these user accounts illuminate alternative conceptualizations of content moderation systems, arguing that an approach to content moderation that places greater priority on user education over punishment may ultimately prove beneficial to platforms struggling to manage content at scale.

Defining content moderation

Throughout this article, I use James Grimmelman's (2015) definition of moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." More specifically, I focus on what Sarah T. Roberts (2014) terms online commercial content moderation: "the organized practice of screening user-generated content (UGC) posted to Internet sites, social media, and other online outlets" (p. 12). Whereas some online communities rely on the work of volunteers and participants in the community to police themselves, commercial content moderation is done for profit by employees or subcontractors.

Often, commercial content moderation relies on the work of outsourced laborers, freelance workers who minute by minute scroll through the worst of the Internet's garbage and make assessments manually as to whether it upholds the community guidelines (Roberts, 2016). Although these laborers are the bedrock of commercial content moderation, they take part in a complex and ever-changing sociotechnical ecosystem. At the moment, users remain critical to the functioning of content moderation: social media companies rely on users flagging content they deem objectionable in order to identify what needs to be removed (Crawford and Gillespie, 2016). Machine learning tools and filters play a role in increasing the efficiency of this process, by scanning the digital fingerprint of an uploaded file to match it to copyrighted content, using Bayesian filters to siphon off spam or skin filters to identify images likely to be pornographic (Roberts, 2014). The role of such tools is likely to grow over time not only because of their increased sophistication but also because of increasing expectations by government regulators that companies remove illegal content, including hate speech and violent extremist content, within predetermined time periods. Despite the likelihood of future regulatory

mandates that will heighten the need for machine learning techniques, at present users are an inextricable element in the moderation process.

Although the examples above suggest commercial content moderation is being subjected to greater scrutiny in the press, it remains a challenging object of study at scale given the black-boxed nature of many social media platforms (Langlois et al., 2009). Moreover, the many layers involved make it challenging to understand content moderation as a complete system: it involves visual interfaces, sociotechnical computational systems, and communication practices. Any attempt at gaining a full view of these operations must take into account their multi-modal effects (Langlois, 2005).

Several researchers have made important initial forays into studying the various layers of content moderation, at the level of discourse (Gillespie, 2010), labor (Roberts, 2014, 2016), policy (Gillespie, 2017), and law (Ammori, 2014; Klonick, 2017). This article seeks to add to this growing body of work by examining commercial content moderation from the perspective of users. In particular, I examine three aspects of users' experience of content moderation processes: user perceptions of how content moderation works, their reports about the impact of content moderation on their lives, and the actions they take (or do not take) in order to seek redress. In doing so, I aim to provide clearer insight into how users are interpreting the obligations of social media companies to the public through the lens of free expression.

Resources for understanding content moderation policy

The technology companies that run social media platforms use a number of different policies to outline their expectations of and relationship to users: terms of service, privacy policies, copyright and fair use policies, advertising policies, data policies, community guidelines or community standards, law enforcement guidelines, developer terms, and payment terms, to name but a few. Although aspects of many of these policies are relevant to users seeking to make sense of content moderation systems, the two that are most salient are the terms of service and community guidelines. Here, I briefly discuss what these policies contain and the information they provide to users seeking to better understand how their content is moderated.

Typically, users will encounter the terms of service of the platform upon joining the platform, and they are required to provide affirmative consent in order to register an account. Terms of Service typically outline the legal terms of the relationship between the company and user, weighing potential precedents, theory, norms, and administrability both in the United States and abroad to develop the rules of speech that govern users. These terms can offer "a sort of jurisprudence of its own" (Ammori, 2014). Often this includes the provision of limitations of liability for each party and establishment of terms of arbitration should the user seek some kind of recourse with the company.

Unless they seek to do so, however, often a user will not encounter the terms of service again as they navigate the site—a common practice is for platforms to include language that states that ongoing use of the platform's services acts as an agreement to be bound by these terms, although they may be changed by the company at any point in time. Terms of Service are not particularly useful to users as an educational tool in general: as they are contracts, they tend to be written in "legalese," although some platforms

have experimented with “translating” the Terms of Service into language more easily interpreted by the average user (Ammori, 2014). Rather, they serve as the legal basis for ongoing user engagement with the platform, outlining users’ rights and limiting the platform’s liability.

Most major social media platforms provide an additional document, frequently termed the Community Guidelines or Community Standards, which establishes in layman’s terms the type of content that is prohibited on the platform. Although there is variation from platform to platform, Community Guidelines frequently outline an overarching vision for the kind of discourse the site seeks to promote and give definition to the kind of discourse it disallows: hate speech, incitements to violence, graphic content, and impersonation, for example.

These terms are more useful for users who seek to understand what sort of content a platform allows: often they provide clear, if somewhat general, operational definitions for categories of content that are disallowed and sometimes include examples to illustrate what can and cannot be posted. These definitions and examples tend to be fairly generic, to allow companies to navigate a core tension in their moderation of content: they have to establish guidelines for a global user base that will be acceptable within the norms and values of the widely variant local markets in which they are present. Acceptable limits on hate speech in Europe or nudity in the Middle East, for example, may be in tension with free speech norms in the United States. The definitions provided, thus, lack precision, which the platforms assert is so that they cannot be “gamed” by individuals who actively seek to violate the rules.

However, companies using commercial content moderation typically maintain separate, non-public documentation that operationalizes the community guidelines at a much more granular level of detail: for example, these documents indicate for content moderators exactly how much blood, gore, or skin necessitates crossing the threshold from acceptable to unacceptable speech (Roberts, 2014). Although the existence of these documents has been confirmed by researchers and some have been published by investigative journalists, they are not made available to the public.¹ Moreover, they change over time in relation to shifting policies and norms within the company, government regulations, and trends in the kinds of content users are posting. Thus, users have only limited insight into the evolving scope of platform content moderation operations at any point in time.

Methods

This article reports the initial findings of OnlineCensorship.org, an advocacy project co-founded by members of the Electronic Frontier Foundation and Visualizing Impact which seeks to shed greater light on content moderation practices by collecting reports from users when their content or accounts are removed from social media sites. The project was formed in 2012 and received a Knight Foundation grant in 2014 before formally launching in November 2015. This article examines data submitted by respondents over the project’s first year, considering how users interpret the role and function of companies in moderating their content, what kinds of content are taken down, and the impact this has on their public expression.

To assess the scope and impact of content moderation on users of social media platforms, we set up a survey at the website OnlineCensorship.org, asking a set of questions about their experience with a content takedown. The launch of the project was publicized to journalists and global advocacy groups via e-mail and received coverage in several mainstream media outlets. A link to the survey was also promoted through social media accounts on Twitter, Facebook, and Instagram. In total, we received 519 responses over an 11-month period stretching from mid-November 2015 to mid-October 2016. Of these responses, 389 related to a content takedown on Facebook, 76 to a content takedown on Twitter, 26 to Instagram, 13 to YouTube, and 6 to Google+. Although we included Flickr as a possible site for content takedowns, we did not receive any reports relating to Flickr during this period. The majority, 221 of the reports, related to an account shutdown or suspension; 105 related to posts that had been taken down, 59 to photos, 18 to pages, and 10 to videos. All other types of content received less than 10 reports, including comments, advertisements, groups, and hashtags.

These results are not generalizable as they represent a group of users who were elected to submit reports. However, they are indicative: they represent the viewpoints of highly motivated users who are likely to be the most concerned about the influence of content moderation. We acknowledged this self-selection bias by focusing the survey on assessing user experiences and perspectives rather than attempting to obtain a sample of the population of content takedowns—information only available to researchers if the companies choose to disclose it.

We anticipated that the subject matter for this project could be attractive to those who might seek to manipulate the data. As such, we took several steps in the survey design to discourage disruption. First, we opted for a page-by-page question design that required respondents to click through multiple pages of content in order to complete a survey. We designed the response fields to offer ample opportunity for respondents to describe their experiences in their own words, rather than simple multiple choice entries that would be easy to replicate. While these measures would not significantly belabor the participation process, they were intended to make it time-intensive for someone to produce falsified results in large quantities. As a final form of verification, we included a page where respondents were offered the opportunity to submit a screenshot as further evidence of the takedown.

Although any survey can be manipulated, the pattern of survey responses gave us a relative degree of confidence that the results were produced organically by respondents and at least did not reflect mass targeting in response to any one particular event or period of media coverage. In the chart below, I provide a longitudinal graph of the timing of survey results. The graph suggests that, following an initial spike upon the launch of the project, there was a relatively constant stream of responses over the course of the year, with the exception of a small peak in responses in March 2016. This uptick in responses coincided with and is most likely explained by a presentation of the project at the Internet Freedom Festival (Figure 1).

We also collected demographic data that can provide some indication of the types of accounts illustrated by the data, which may be useful context for interpreting any bias or skew in the sample: of the 510 respondents, 295 were based in the United States, the largest country represented in the sample. As no other country had this level of concentration, I will represent the geographic distribution of the remainder of the sample by

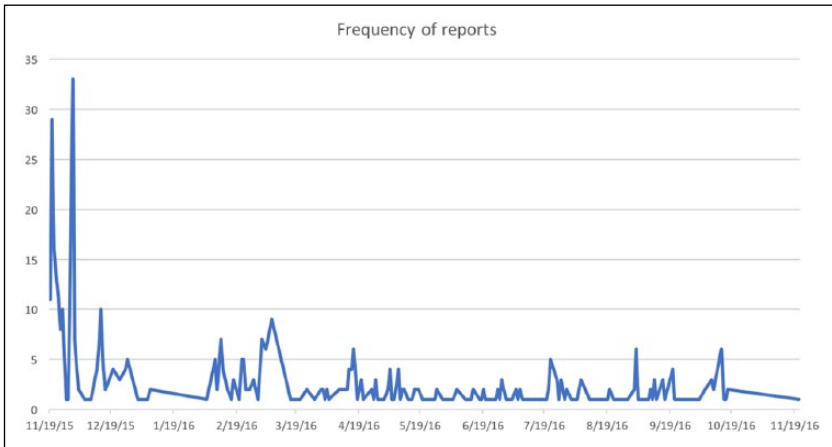


Figure 1. Frequency of reports submitted to OnlineCensorship.org.

region: 4 in Sub-Saharan Africa, 47 in North America (not including the United States), 9 in East and Southeast Asia, 1 in Central Asia, 94 in Europe, 12 in Latin America, 7 in the Middle East and North Africa, 18 in Oceania, and 11 in South Asia. Twelve respondents elected not to provide their place of origin. There was a similarly high concentration in language; 444 reports out of the content in question related to English language posts, while 16 related to German language posts, 9 to Spanish language posts, and 9 to Portuguese language posts, with the remainder relating to only a handful of cases in other languages.

A final issue of note is that although the users submitting reports come from a wide range of countries, the companies they discussed are all headquartered in the United States. Content moderation practices for non-US-based countries may vary widely: for example, social media companies in China have a mandate to monitor social media content that US companies are protected from under Section 230 of the Communication Decency Act, an important legal distinction with downstream effects on content policy and moderation practices. In this article, I focus on the users of a subset of US-based social networks, which may not be representative of the experiences of the users of social networks subject to different laws and that engage in different content moderation practices.

Folk theories about content moderation

Many social network users develop “folk theories” about how platforms work: in the absence of authoritative explanations, they strive to make sense of content moderation processes by drawing connections between related phenomena, developing non-authoritative conceptions of why and how their content was removed (Eslami et al., 2015; Kempton, 1986). These theories are often incomplete, are generated out of each user’s personal experiences, and serve to help them both to filter information and to make decisions about how they interact with the world (Gelman and Legare, 2011; Jones et al., 2011; Rader and Gray, 2015).

Although the study surfaced a number of folk theories about how and why their content was taken down, users' interpretations tended to pinpoint human intervention as the primary cause. The most common theory offered by users was that their content was flagged by another user. Several users thought the flagger was someone specific that they already knew: "My friend got mad at me," said one user (User report, 6 October 2016), while another user said their account was shut down "Because I was 100% targeted" (User report, 1 April 2016). Often, the users noted they were flagged in the midst of a contentious discussion, whether about politics, religion, or interpersonal relationships, and attributed the flag to the party they were in conflict with. For example, one Facebook user reported that "Facebook is allowing someone to use the report feature to harass me and it isn't fair" (6 March 2016), while another said that "this is about a small group of people perpetuating myths (IMO) that are censoring me using FB's overwhelmed systems while they bully, threaten and harass me" (31 May 2016).

In other cases, the users perceived the source of the flag as a disembodied "they," perceiving that some unknown party opposed their speech. Although in some cases the users attributed this to a personal vendetta, in other instances they related the flag to their group identity, perhaps drawing connections between their personal experiences and many social media platforms' well-documented challenges policing hate speech (European Commission, 2018). For example, one Facebook user said, "Someone who does not like Muslims is using the report feature to harass and bully me and Facebook is allowing it to happen" (User report, 6 March 2016). Another said, "In light of increased violence and discrimination against transgender people, my Facebook profile has been getting attacked by transphobes who reported my picture" (User report, 18 April 2016).

Several users attributed the reason for the content takedown to perceived political bias on the part of the company. For example, one user said, "I suspect that facebook [sic] has one or more male mods who are hostile to feminists and minorities in general" (User report, 16 April 2016). As the data were collected in a year running up to the US presidential election, many of these users imagined that the companies sought to influence election results through censorship of social media content. "I think it was censored because either Twitter management is protecting Clinton, or her campaign has called for censorship. Either way it's wrong," one user said (User report, 2 July 2016). Although US politics figured prominently due to a heavy representation of US-based users in the sample, other forms of political bias suggested included perceived discrimination against women, Muslims, and both pro- and anti-Zionist sentiments.

Finally, some users described expectations of more stringent legal responsibilities for platforms to uphold norms of free speech under US law (a theme also likely to be reflective of the heavy prominence of US-based users in the reports). Often, these users expressed the belief that, regardless of the reason for the takedown, social media companies *should* be held to the same legal requirements as the US government under the First Amendment: "I believe the content was censored because Facebook does not protect free speech or uphold the First Amendment to the US Constitution, the Bill of Rights, right of free speech and free expression," said one user (User report, 24 March 2016). Another user said,

I don't feel it is within the scope of social media sites to regulate or censor topics, especially unpopular topics. It sets a negative expectation and legal precedent for services to become

responsible for the content their users post. Facebook has opened a pandora's [sic] box by banning legal but unpopular gun groups. Which unpopular group will they ban next? (User report, 14 April 2016)

These expectations, though strongly expressed, are unsupported by current regulations of social media platforms: Section 230 of the Communications Decency Act acts as a "safe harbor" provision, relieving social media companies of the obligation to moderate most forms of content even if they elect to do so voluntarily.

Overwhelmingly, the users in this study attributed purposeful targeting by other human actors as the primary reason for the removal of their content, rather than seeing algorithms or other forms of technological intervention as the cause. This finding is resonant with other studies of users' folk theories about how social networks operate: for example, in their analysis of user folk theories about Facebook's newsfeed, Eslami et al. (2015) found that some users tended to over-attribute actions of Facebook's curation system to their own family and friends, much like some users attributed the reason for the removal to other users over other, non-human platform activities. In another study of folk theories about Facebook's algorithm, Eslami et al. (2015) found that one common theory users held was that Facebook was a "powerful, perceptive, and ultimately unknowable" entity and that some users believed the company actively suppressed stories about religion and politics. Several of the users in this study similarly attributed content removals to active intervention by the company, although they ascribed a more overtly political role to Facebook and Twitter than the users in the Eslami et al. (2015) study.

The majority of user posts on most social media platforms are indeed removed as a result of flagging by other users. But there are a range of reasons why a user might flag another user's content: a desire to take part in upholding community norms and values, to remove content they see as offensive, to intervene to protect another user, or, indeed, to target and silence another user who they dislike. Moreover, there are non-human platform activities that could also have been responsible for the removal of content: spam filters that analyze users' behavior patterns or hashing algorithms that remove copyrighted content, child pornography, and extremist content upon upload. Some of the kinds of activities users reported did not, in fact, even involve the active removal of their content: instead, they suspected they were "shadowbanned," where their content is made invisible to other users without actually being removed entirely. Neither the users, nor the platform, nor this author are well placed to accurately assess why these users' content was flagged—these assessments can only, at best, serve as interpretations of possible motivations.

Impact of content moderation on users

The accounts presented by respondents illustrated that content moderation produced a range of impacts on their lives that transcended online and offline experiences. Often, content moderation is described as a form of censorship or restraint on a user's voice; it quite literally removes the content of their speech, and in the case of an account suspension prevents their access to a channel for future expression. Although many users did discuss their experience with content moderation as an inhibition of their capacity for

self-expression, the accounts surfaced a wider spectrum of consequences, some of which were particularly detrimental to users who are already in a marginal position in society.

The most vivid accounts came from users who had experienced account suspensions or bans, the most stringent form of moderation. Bans on access to one social media platform often led to the inability to access other platforms: since services like Spotify, the Huffington Post, and Tinder often use social logins, limits to accessing one platform can mean the user is no longer able to access a range of other services. Other forms of loss reported by users range from the relatively mundane (such as the inability to participate in contests and drawings) to the deeply intimate (losing access to stored photos of family members).

Many users lamented losing the ability to communicate with others: one user reported not knowing that a family member had passed away, as the social media platform they were restricted from was the primary means of communication within their family network. The accounts foreground the importance of social media platforms as a mechanism for staying in touch with support networks: for these users, losing access does not mean solely the loss of a platform for their speech, but the loss of an essential channel of communication with the outside world and likely prospect of increased social isolation. For example, one user responded, “I know it’s crazy but I’m lonely without Facebook” (User report, 4 June 2016), while another user who was living in another country than their spouse said, “Even my own wife uses Facebook constantly, so I am more left out than less immediate family and friends” (User report, 6 May 2016). These effects were particularly impactful for the elderly and disabled: one user said, “I worry that if I am ever dying & attempt to reach out to my friends, Facebook will block me from contacting them, which frightens me & causes me considerable angst” (User report, 16 January 2016). Another said,

I am disabled and very sick ... I have had my tagging ability taken away three times and my ability to post taken away once. If I cannot tag, I cannot guarantee that I am alerting anyone! If I can’t post, I am really in trouble. (User report, 6 January, 2016)

Several users reported that the suspension of their accounts had a negative impact on their professional life, often causing irreparable damage for those who rely on data analytics and ad revenue to support them financially. For example, one user said, “My page is my life—3 years of hard work an committment [sic] to build a community online for women, my page was organically reaching on average 3.5 million a week” (User report, 9 July 2016). In some cases, an erroneous takedown meant that a user’s followers were completely wiped from their account upon restoration, requiring the user to rebuild their community of followers from the ground up.

Artists in particular often reported that social media platforms such as Instagram serve a critical role as the primary mechanism they use to obtain referrals for business—thus the shutdown of their accounts had tangible economic impact. In one case, a photographer reported explicitly seeking to avoid the violation of community guidelines on Facebook by posting a link to a story about a portrait series he had worked on. The link auto-generated an image that contained nudity, one of 2 out of 12 possible images that could have been selected from the series, most of which did not contain nudity. Not being

able to select which image was displayed, the photographer's account was suspended, preventing him from sharing new work and connecting with potential clients (User report, 1 April 2016). In other cases, losing access to an account had downstream effects for users who could no longer administer pages; for example, one alternative news site was shut down for a full month when both of its admins had their accounts suspended.

These accounts suggest that content moderation can have wide-ranging impacts on users' experiences that far exceed their capacity for self-expression. Although for some punitive measures serve only as an annoyance, for others the impact is substantial. Many users rely heavily on particular social media platforms as a primary channel of communication with friends and family, as professional networks, or as a means of accessing other media platforms. When they lose access to these channels, they do not have the option of simply going elsewhere—they have lost their community in addition to their access to the platform. These respondents expressed a desire that social media companies acknowledge the ways in which their platforms provide a public good: that they support systems of communication that are deeply interwoven with social, political, and economic life.

Affective dimensions of content moderation

Although the users who submitted reports to OnlineCensorship.org ranged widely in their perceptions of how content moderation worked, many of them expressed confusion and frustration about the process and suggested that content moderation systems appeared to be designed to escalate these emotional responses. As the accounts below will illustrate, because content moderation systems do not offer users much opportunity for human interaction, participating in the survey itself seemed to offer respondents a venue to vent about their feelings. In fact, some users returned to submit increasingly heated follow-up reports after multiple suspensions, as the use of all caps indicates in this example:

Well, I'm back with YET ANOTHER account ban. I know for a fact I DID NOT spam any comments, I DID NOT put a bot on my account, I ONLY HAD IT FOR A FEW DAYS AND THIS HAPPENED. (User report, 1 April 2016)

Users generally expressed uncertainty as to why the company thought they had violated their policies and how their content was taken down, even if they said they had examined the community guidelines. For example, one user whose account was suspended from Facebook for violation of the company's "inauthentic name" policy said,

It is my account that has been suspended/terminated (the way FB Authorities call it as "deactivated") without any prior notice. It's been over 8 months by now. I took action and contacted FB for a proper explanation, apology and reactivation of my account twice already ... all they did both times was copy/pasting me the same cliché regarding general FB policies of what might have been the cause of such action on their end. Nothing even personalized about the specific problem/solution regarding "my" account; just generic "possible causes"! (User report, 16 March 2016)

A frequent complaint from users was that the company did not disclose to them specific details about how they violated the community guidelines, for example, by not telling them

which policy they violated. This confusion led to repeat offenses, as this report by a Facebook user suggests:

I honestly actually have no idea what, who, or why I've been banned this time ... This time I have so far sent 4 appeals simply asking what reason I've been banned and as of yet have still gotten no response from Facebook! ... Like I said, all I want is an explanation so that if nothing else I can try to avoid this in the future! (User report, 4 May 2016)

The perceived absence of a real person on the other side of the computer screen was a particular source of frustration for many users: one person said, "I've been blocked 3 times in march [sic] for not respecting the community guidelines and I feel like they're becoming stricter but Facebook won't provide me with any human interaction or explanation" (User report, 2 April 2016). Still others suggested that even when they did reach a real person, they found little additional information that would be helpful to them in navigating content policies: in the words of one user,

After I appealed I received an email from someone called Ron at Facebook's Pages Support section saying "I'm here to help." I emailed Ron explaining that I didn't understand why the page had been unpublished and I asked him to say which post contained (as they claimed) "malicious or misleading content." I offered to comply with facebook's [sic] wishes and delete any post they thought was a problem. His reply didn't provide any detail at all, it simply said "We have a no-tolerance policy concerning this infraction and your page is ineligible to be republished." I still don't know which post triggered the censorship. (User report, 21 July 2016)

These findings are consonant with a broader trend in both digital and physical service contexts toward adoption of self-service technologies—interfaces in which customers have minimal to no contact with human representatives (Ostrom et al., 2002). Although self-service technologies may increase firms' efficiency, customers have consistently expressed a preference for human interaction, often because self-service channels are poorly designed (Immonen et al., 2018). Those who are more technically savvy are slightly more likely to see such channels positively (Immonen et al., 2018; Meuter et al., 2000), whereas those who have higher levels of technology anxiety seek to avoid them entirely (Meuter et al., 2003). Unsurprisingly, individual attention has a significant impact on customer satisfaction (Babbar and Koufteros, 2008).

The use of automation and limited opportunity for human interaction in content moderation systems likely served to increase users' frustration with the process and motivated them to seek out other channels for information, as in the case of the user who contacted the Support team. This was illustrated in the tone several users adopted in the survey responses themselves—their grievances were palpable as they filled out yet another form in hopes that it would have some effect on the company's response. Frustration was a distinctive, if unsurprising, dimension of all the responses.

Users' feelings of confusion were another dominant affective response: by and large, those who interact with content moderation systems report they were unable to learn what triggered the content removal or account suspension. Many of these users expressed

a desire to understand why this happened so that they could avoid the experience in the future: although it is likely that they would prefer human interaction regardless, clearer messaging may suffice to reduce these feelings of confusion.

Due process, appeals, and channels for user agency

Most (though not all) platforms considered in the survey offer users some form of redress when they feel their content is taken down in error. The form this takes may differ from platform to platform: for example, while Facebook and Instagram only allow users to appeal the suspension of their account, Twitter, Google+, and YouTube allow appeals at more granular levels, including posts and videos. In most instances, users are notified about the mechanisms for appeal available to them through in-app notifications at the time the action is taken against them.

The reports suggest that many users would like to take advantage of the opportunity to appeal, and that in many cases users believe their content was taken down in error (in some cases we were able to verify whether the content was in violation through screenshots provided by users).² Roughly half, 230, of those reporting said they had appealed the removal of their content. However, many of these respondents reported running into problems during the appeals process. One user who had attempted to appeal said of the process, “They don’t have one. I contacted them to complain and got no response” (User report, 13 April 2016). Another said, “We did receive a response on the appeal, but it was just a reiteration of the earlier decision” (User report, 14 April 2016).

Other users said that although they were notified of the ability to appeal, they were unclear on what to do next: “Facebook pops up a notification saying your group has been disabled, and that an appeal must be made within 7 days or the group will be permanently deleted. There was no link or instructions to begin the appeal process,” reported one user (User report, 14 April 2016). Another said, “There is no real ‘appeal’ process with Facebook. You can click on their ridiculous series of ‘emojis’ and leave a Comment, but that’s it. You might as well be speaking into a void” (User report, 14 December 2016).

As in the notification process, users expressed frustration that they were not offered an opportunity for human interaction in seeking redress for the removal of their content. This frustration was often reinforced when the appeals process did not result in a resolution of the problem: in many cases, appealing a content removal results in the content going through the same review process and thus does not incur any escalation or additional level of scrutiny for the content. Furthermore, as one user reported, even the successful resolution of an appeal was no guarantee that they would not be suspended again in the future:

My birth given name is [redacted] Pizza. Every year or so they make me scan documents and prove that Pizza is a real name. This year they locked me out of my account. It’s annoying because I’ve sent the documents three times now and the messaging function has become important being an entertainer. I had to get friends to send messages on my behalf which is very unprofessional. (User report, 20 September 2016)

Roughly half, 245, of those reporting said they did not appeal the removal of their content. Of these, 100 said they did not appeal because they did not know how to, and 60

of them said they did not expect a response. Some of these users nevertheless attempted to seek redress from the company via other means. Several sought to contact the company using accounts on other social media platforms: for example, a user who had their content taken down from Instagram tried tweeting the company's account on Twitter. Another Instagram user tried to use the "report a problem" feature in the app, receiving a message back saying the company was happy they are helping to improve Instagram (User report, 21 September 2016).

Other users turned to other locations on the platform where they felt they would be more likely to get a human response. For example, one user said, "After I've [sic] got banned I created a support ticket, which doesn't even appear in the support console. I asked why they don't stick to their own community guidelines and why they ban my paintings" (User report, 10 May 2016). Another user said they were able to get a response by posing to the advertiser page, but that ultimately this did not result in the restoration of their content (User report, 13 June 2016).

These accounts from users suggest the importance of due process and systems for redress not only to the effective functioning of content moderation systems but also to developing and maintaining a positive affective relationship between platforms and their users. The appeals process as currently designed functions as a kind of bureaucratic limbo, giving users an outlet that channels their frustrations through some form of action in response to the takedown, but most often serves to exhaust them rather than offer a meaningful outlet for redress. Moreover, the absence of human interaction heightened frustration for some users, as poignantly expressed by the respondent who likened the appeals process to shouting into a void. Despite this, some users sought out a variety of inventive mechanisms to assert their agency, turning to various parts of the platform in order to break through the systems of automation.

Changes that took place over the course of this study in the content moderation process for two platforms, Facebook and Twitter, are further indicative of the companies' orientation toward the appeals process. Both platforms began experimenting with a new approach to content moderation that, instead of making permanent decisions about the suspension of accounts, would give them temporary suspensions of 12 hours up to 30 days, effectively placing users in "time out" for violating community guidelines. The length of these temporary bans varied depending on the severity of the infraction and whether the user had been penalized for an offense in the past, but the companies provided no further detail on which kinds of violations were matched to more severe punishments. Most notably, temporary suspensions do not allow users to appeal the platform's decision: instead, they have to wait out their punishment and will eventually be allowed back online.

From the perspective of a platform experiencing growing pains, such a system would appear to solve real challenges: automatic bans inhibit users from engaging in harmful behavior and cut down the overall amount of content for moderators to review by eliminating the system for appeals. However, if the overall objective of a content moderation system is to encourage better behavior on the part of users, such a system doubly fails: not only does it not address the need to educate users about the reason they violated a content policy, it offers them no opportunity for engagement with the platform to learn. By eliminating due process, automatic bans presume that users both intended to break the rules and are thus unable to learn how to do better.

Perhaps this is, in fact, the case for many social media users: the responses to this study tended to make up the “gray area” of cases—instances where the content that was removed was not wildly in violation of the guidelines and required some interpretation. Their accounts suggest a user base that is emotionally engaged in participating in the life of a social network, invested in learning from mistakes, and confused about where things went wrong. In short, they are exactly the kinds of users who make up the kind of “town square,” “global village,” or “community” that these platforms themselves say they seek to cultivate—but current content moderation systems do not give them much opportunity to participate or grow as citizens of these spaces.

Conclusion

This project was designed to shed light on user experiences with content moderation and, as might have been predicted, illustrated a number of sources of frustration: confusing interfaces and messaging that did not show users where they went wrong, appeals processes that seemed to go nowhere, and minimal opportunity for users to interact directly with employees of the platform. This led users to develop their own folk theories about how platform moderation works that placed blame on human intervention as the primary cause for the removal of their content, rather than the broad and complex range of socio-technical factors that make up content moderation systems. Analyzing these findings raised a question that is perhaps more fundamental than I set out to examine at the outset: what is it, exactly, that content moderation systems are designed to accomplish?

As they currently are designed, content moderation systems remove content at massive levels of scale, but do not do much to educate users about where they went wrong. Although in large part social media platforms are designed to make their users feel good, to want to engage and share with others, the design of content moderation systems works at cross purposes with this intention: they make people feel confused, frustrated, and as though they are “shouting into a void.” The overwhelming reliance on flagging mechanisms to identify content to be removed reinforces a sense among users that platforms are a place where they can be targeted for their speech, beliefs, or identity, and there are ample examples that indicate that flagging can be used for just that purpose. The absence of an effective appeals system means that social networks feel dehumanizing for users, and they are left without any recourse when they make a mistake or are penalized in error.

More positively, the frequent appeals made by platforms to the rhetoric of community and to the notion of social networks as a public sphere did indeed resonate among these users. Some of them expressed a desire to change their behavior, to learn so that in the future they will no longer risk their accounts being shut down. Others sought opportunity to take action as citizens, to have a voice in shaping the policies set by companies through participation and, on occasion, through acts of civil disobedience. But although many of the users who participated in this study showed every intent of acting as a member of a community—admittedly, not all, but a substantial number—these systems as currently designed do not offer much opportunity for them to do so.

There are no easy solutions: some of the kinds of interventions that these users are calling for would be extremely resource-intensive to deploy at scale, and their differing ideals of what kinds of content platforms should allow may ultimately be

irreconcilable. But it may be worth considering what a content moderation system designed to educate and engage with users as community members might look like: simple design interventions like providing more detail and clarity about content moderation processes, or including explanations of the policy violated when notifying a user of a content takedown, would go a long way toward encouraging users to better police themselves—and, importantly, this could ultimately mean *reducing* the amount of content platforms have to moderate, rather than increasing it. Developing a different standpoint from which to position the “users” of social media platforms could be a generative space for content moderation systems to develop and, ultimately, could point to alternative ways of thinking about what content moderation and indeed platforms themselves are designed for.

Acknowledgements

The author wishes to thank the OnlineCensorship.org team: Jillian York, Ramzi Jaber, Jessica Anderson, Kim Carlson, and Matthew Stender, as well as the reviewers for their useful feedback during the revision process.


Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a fellowship provided by the Annenberg School for Communication and Journalism and the USC Graduate School.

Notes

1. For examples, see Hopkins (2017) and Angwin and Grassegger (2017).
2. The screenshots proved to be an imperfect verification system; not all users submitted screenshots, and not all screenshots gave us enough context to evaluate their reports. For privacy reasons, we did not require screenshots as part of the submission system.

ORCID iD

Sarah Myers West  <https://orcid.org/0000-0001-5793-2405>

References

- Ammori M (2014) The “new” New York Times: free speech lawyering in the age of Google and Twitter. *Harvard Law Review* 127(8): 2259.
- Angwin J 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>
- Babbar S and Koufteros X (2008) The human element in airline service quality: contact personnel and the customer. *International Journal of Operations & Production Management* 28(9): 804–830.
- Baker K (2013) Twitter CEO Costolo focused on “building global town square.” *Bloomberg*. Available at: <https://www.bloomberg.com/news/articles/2013-03-25/twitter-ceo-costolo-focused-on-building-global-town-square->
- Crawford K and Gillespie T (2016) What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18(3): 410–428.

- Eslami M, Rickman A, Vaccaro K, et al. (2015) "I always assumed that I wasn't really that close to [her]": reasoning about invisible algorithms in news feeds. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems (CHI'15)*, Seoul, Republic of Korea, 18–23 April.
- European Commission (2018) Commission recommendation on measures to effectively tackle illegal content online. *European Commission*. Available at: <https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online>
- Facebook (2016) Community standards. *Facebook*. Available at: <https://www.facebook.com/communitystandards>
- Gelman SA and Legare CH (2011) Concepts and folk theories. *Annual Review of Anthropology* 40(1): 379–398.
- Gillespie T (2010) The politics of "platforms." *New Media & Society* 12(3): 347–364.
- Gillespie T (2017) Governance of and by platforms. In: Burgess J, Poell T and Marwick A (eds) *The SAGE Handbook of Social Media*. New York: SAGE.
- Grimmelmann J (2015) The virtues of moderation. *Yale Journal of Law and Technology* 17(42): 42–109.
- Halliday J (2012) Twitter's Tony Wang: "we are the free speech wing of the free speech party." *The Guardian*. Available at: <https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech>
- Hopkins N (2017) Revealed: Facebook's internal rulebook on sex, terrorism and violence. *The Guardian*, 21 May Available at: <https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence>
- Immonen M, Sintonen S and Koivuniemi J (2018) The value of human interaction in service channels. *Computers in Human Behavior* 78: 316–325.
- Jones NA, Ross H, Lynam T, et al. (2011) Mental models: an interdisciplinary synthesis of theory and methods. *Ecology and Society* 16(1): 46.
- Kempton W (1986) Two theories of home heat control. *Cognitive Science* 10(1): 75–90.
- Klonick K (2017) The new governors: the people rules, and processes governing online speech. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2937985
- Langlois G (2005) Networks and layers: technocultural encodings of the World Wide Web. *Canadian Journal of Communication* 30(4): 565–583.
- Langlois G, Elmer G, McKelvey F, et al. (2009) Networked publics: the double articulation of code and politics on Facebook. *Canadian Journal of Communication* 34: 415–434.
- Meuter ML, Ostrom AL, Bitner MJ, et al. (2003) The influence of technology anxiety on consumer use and experiences with self-service technologies. *Journal of Business Research* 56(11): 899–906.
- Meuter ML, Ostrom AL, Roundtree R, et al. (2000) Self-service technologies: understanding customer satisfaction with technology-based encounters. *Journal of Marketing* 64(3): 50–64.
- Ostrom AL, Bitner MJ and Meuter ML (2002) Self-service technologies. In: Rust RT and Kannan PK (eds) *E-Service New Directions in Theory and Practice*. Abingdon: Routledge. Available at: <https://www.taylorfrancis.com/books/e/9781315291284/chapters/10.4324%2F9781315291291-10>
- Rader E and Gray R (2015) Understanding user beliefs about algorithmic curation in the Facebook news feed. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems (CHI'15)*, Seoul, Republic of Korea, 18–23 April.
- Roberts ST (2014) *Behind the screen: the hidden digital labor of commercial content moderation*. PhD Thesis, University of Illinois, Chicago, IL.

Roberts ST (2016) Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste. *Wi: Journal of Mobile Media* 10(1): 1–18.

Author biography

Sarah Myers West is a doctoral candidate at the Annenberg School for communication and journalism, University of Southern California and an affiliate researcher at the Berkman-Klein Center for internet and society at Harvard Law School. Her research examines the role of technology companies in governing speech and new formations of networked public spaces. Her work has been published in venues such as *Policy & Internet*, *Business and Society*, and *International Communication Gazette*.