Received: 2 June 2023 | Accepted: 12 July 2023

DOI: 10.1002/poi3.359

RESEARCH ARTICLE

The weaponization of platform governance: Mass reporting and algorithmic punishments in the creator economy

Colten Meisner

Department of Communication, Cornell University, Ithaca, New York, USA

Correspondence
Colten Meisner, Department of Communication, Cornell University, 498 Mann Library Bldg, Ithaca, NY 14853, USA.
Email: ccm252@cornell.edu

Funding information
Cornell Center for Social Sciences

Abstract

Amid wider discussions of online harassment on social media platforms, recent research has turned to the experiences of social media creators whose compulsory visibility renders them vulnerable to frequent attacks, ranging from persistent trolling to much more insidious, offline acts of violence. This study examines a contemporary form of harassment against social media creators known as "mass reporting," or the targeted, automated flagging of creators' online content to trigger content takedowns and account bans. Mass reporting is particularly challenging for social media creators because of its weaponization of platform infrastructures for community governance, leaving creators with few avenues of support after orchestrated attacks and restricting their access to platform support. Drawing on in-depth interviews with social media creators who have been subjected to mass reporting on their content, this study builds an understanding of the ways in which tools for platform governance, such as content reporting, can be weaponized to harass and introduce vulnerabilities for creators.

KEYWORDS

creators, flagging, harassment, influencers, mass reporting, platform governance

INTRODUCTION

Frustrated with the perceived unevenness of automated content moderation on TikTok, a Danish teenager known only by his online username "H4xton" created the TikTok Reportation Bot: a computational tool comprised of sets of code, made publicly available on GitHub, that sends content violation reports en masse to TikTok with the aim of having content removed or the post's creator banned. The bot accomplishes content takedowns by flooding TikTok's reporting system with repeated flags that will trigger automated content moderation. In an interview with the *Los Angeles Times*, H4xton understood the tool as a force for good on TikTok. Explaining his motivations for creating and sharing the bot, he said, "I want to eliminate those who spread false information or … made fun of others" (Contreras, 2021). Since its launch, H4xton's tool has been shared widely online to mobilize against—ostensibly harmful—content on TikTok. But however virtuous his intentions may have been, his technical expertise has been publicly circulated to those with goals far more insidious.

This phenomenon—which many creators refer to as "mass reporting"—is not an entirely new challenge for platform governance, and it occurs across many major social media platforms, including Facebook and Instagram (Kayser-Bril, 2021). While advocates contend that mass reporting can limit the visibility of rule-violating content, the practice provides a blueprint that has been harnessed for more insidious purposes, including networked harassment and targeted content takedowns against already marginalized creators (Clark-Flory, 2019). Coordinated abuses of reporting systems—such as groups of users organizing together to manually report users or posts—have posed challenges for platform governance since the early days of Craigslist (White, 2012), but the automation of content violation reporting through tools like H4xton's TikTok Reportation Bot magnifies these problems for platform companies and their online communities. Mass reporting can be frustrating for casual users of social media platforms, but for full-time social media creators, the threat of content takedowns, account suspensions and bans, and reductions in visibility through algorithmic recommendations can have severe and lasting consequences for their professional images, financial livelihoods, and emotional well-being (Are, 2022; Bishop, 2019; Cotter, 2021; Duffy & Meisner, 2023). Research on content takedowns in the creator economy rightly addresses unfairness in automated content moderation (e.g., algorithmic reduction; see Gillespie, 2022), though considerably less is known about the experiences of creators attempting to manage abuses of the infrastructures for platform governance, such as the role of user-driven flagging in content reporting and moderation (Crawford & Gillespie, 2016).

This study therefore examines how a specific group of social media users—content creators—navigate their encounters with platform governance at a time when the infrastructures for policing harassment and incivility have, ironically, become weaponized against them. Drawing on in-depth interviews with 18 social media creators on Instagram, TikTok, Twitch, and YouTube who experienced mass reporting, this study's findings illuminate a portrait of social media governance that has been usurped by motivated, networked actors in an attempt to render targeted creators—and their content—invisible. I document three dimensions of the weaponization of platform governance in the case of mass reporting: manipulating reporting infrastructures, removing posts and punishing creators, and reifying platform opacity and precarity. Taken together, these creators' accounts broaden our understanding of content takedowns enacted through coordinated harassment at a time when information and power asymmetries about platform governance continue to intensify between platform companies and social media creators.

CONTENT MODERATION ON SOCIAL MEDIA PLATFORMS

Platform companies have not historically embraced their roles as media companies—that is, as regulators of the content published on their sites—but instead as tech intermediaries merely hosting user-generated content without liability (Gillespie, 2010; Napoli & Caplan, 2017). Research on platform governance captures "the growing body of work addressing the political effects of digital platforms (governance by platforms), as well as the complex challenges that the governance of platform companies presents" (Gorwa, 2019, p. 855). Governance *by* platforms includes the massive content moderation apparatuses, outsourced in large part to contracting firms, that are responsible for detecting and reviewing content uploaded to social media platforms that violate the platform's community guidelines (Gillespie, 2018; Roberts, 2019). Both academic and industry reports tend to disproportionately frame the outcomes of content moderation as a process of content *removal* instead of the more common *reduction* in visibility, wherein content remains available on a platform but must be manually located rather than algorithmically curated for users (Gillespie, 2022; Zeng & Kaye, 2022).

Content moderation is not experienced evenly across all users, in part because algorithmic content moderation is trained on data that "may be biased in ways that are worth investigating before implementing such systems" (Binns et al., 2017, p. 8). Specifically, three groups of social media users—transgender people, Black people, and political conservatives—were found to experience content moderation more frequently than others but for distinct reasons tied to discussions of personal identities or that represent dominant cultures of use within each group that attracts content moderation (Haimson et al., 2021). These tensions between particular groups of users and the platform ownership resonate with what Sybert (2022) characterizes as *contested* platform governance, in which a clash of values between platform and user interests manifests in challenges to content policies and site features.

Content flagging is one the most ubiquitous forms of content moderation, beginning in early online communities and evolving with the development of contemporary social media platforms. Flags are communicative tools that can reflect a wide variety of motivations, with some cases closer to the intended usage than others. For instance, flags might be seen as a "prank between friends, as part of a skirmish between professional competitors, as retribution for a social offense that happened elsewhere, or as part of a campaign of bullying or harassment—and it is often impossible to tell the difference" (Crawford & Gillespie, 2016, p. 420). The varied use of flags also reflects the wide-ranging nature of flaggers: some users flag to protect their communities from toxicity, while others frivolously flag content as disruptors in communities (Kou & Gui, 2021). Some platforms, such as Craigslist and Reddit, prioritize community governance and depend considerably on "good citizen" labor, which includes encouraging flagging efforts within subcommunities that develop their own norms and rules (Matias, 2019; White, 2012). Yet, even in early instances of community-driven flagging, there were signs of coordinated manipulations of content reporting features similar to challenges with mass reporting seen across today's platforms. Such organized flagging efforts are, as Gillespie (2018) put it, "a kind of 'user-generated warfare'—all through the tiny fulcrum that is the flag" (p. 93).

GOVERNING COMMUNITIES IN THE CREATOR ECONOMY

The commercialization of social media ushered in the prominence of social media creators as profitable cultural producers for platform companies. The labor of social media creators is most aptly understood as a labor of visibility that involves the work of maintaining a public-facing self to capture attention in platform communities (Abidin, 2016). Depending on the affordances of platforms where creators produce their content, they are expected to navigate not only the production of satisfying content for their audiences but also managing market, industry, and algorithmic precarity (Duffy et al., 2021), as well as the affective labor of maintaining some degree of audience community through sustained relationship development over time (Uttarapong et al., 2021).

Because social media creators are their own producers *and* promoters, they often circulate communally and socially informed folk theories about algorithmic recommendations ("algorithmic gossip"; Bishop, 2019) ranging from intentionally producing content that will succeed in reaching wide audiences to creatively misspelling words when discussing controversial topics to avoid automated detection systems. For creators producing content in stigmatized genres—such as sex workers, cosplay performers, and many other categories of creative labor—the widespread application of automated content moderation frequently and disproportionately flags content from already marginalized creators in initial reviews, without regard to the creators' attention to formal content policies while producing content for (public) platform audiences (Are, 2022; Duffy & Meisner, 2023). Adding insult to injury, many marginalized creators experience difficulties *appealing* initial, automated platform decisions, which marginalized creators attribute to their relative lack of visibility compared to widely popular streamers (Thach et al., 2022). But importantly, the inscrutable nature of platform algorithms creates the necessity for algorithmic gossip and folk theories about visibility in the first place. In Cotter's (2021) analysis of Instagram's "shadowbanning" controversy, she suggests that the debates over the existence of shadowbans—or, more specifically, visibility moderation—draw attention to an unmatched epistemic contest about algorithmic visibility in which platform companies will *always* have the upper hand over creators and users. This power asymmetry makes space for content removals and deplatforming but not for educating users about policy-violating infractions (West, 2018) or recovering and rehabilitating accounts (Are, 2023).

In addition to removing content that violates platform policies, platform governance also serves an essential role in mitigating toxicity and harassment in online communities. In online communities that emphasize community-led and volunteer moderation, governance can support the development of both positive and negative norms, with community rules on sites like Reddit helping to reinforce "toxic technocultures" in some subcommunities (Massanari, 2017). In other communities, social media creators are commonly subjected to hostility, with one in three creators claiming harassment is a regular occurrence in their work (Thomas et al., 2022). This study conceptualizes mass reporting not only as a manipulation of infrastructures for platform governance but also as a form of networked harassment (Marwick, 2021; Marwick & Caplan, 2018) invoked to target—and silence—social media creators, often those identifying with communities already subjected to marginalization. In doing so, I explore how this case of creator harassment through reporting infrastructures can build our understanding of the potential of networked users to challenge or reproduce platform governance.

METHODS

I conducted 18 in-depth interviews with social media creators producing content across social media platforms including Instagram, TikTok, Twitch, and YouTube who had experienced mass reporting as a form of coordinated, networked harassment against their account and content. Interviews primarily focused on how creators identified the threat of mass reporting, managed platform punishments, and resisted or responded to mass reporting. Many participants in this study were recruited as part of a larger project about algorithmic discrimination on social media, and thus, the creators represented in this study overwhelmingly identified with historically marginalized identities and/or produced content stigmatized in the creator economy. Participants responded to interview solicitations on social media platforms seeking creators who had experienced various forms of harassment, including mass reporting. Creators were compensated with a US$25 stipend for their participation in a single 30 to 45 minute interview. In addition to speaking with creators directly about their experiences through interviews, I also worked with a research assistant to analyze the community guidelines of Instagram, TikTok, Twitch, and YouTube and their parent organizations, if applicable, to understand platform rhetoric and positioning around user-generated flagging and reporting mechanisms as part of their overall content moderation apparatus.

I adopted a grounded theory approach (Glaser & Strauss, 1967) and iteratively refined the interview protocol throughout data collection as interviews were conducted, adding new questions and revising others based on the responses and experiences shared by interviewees. The interviews were pseudonymized and transcribed by a professional service. The initial coding round identified substantive and descriptive themes recurring across interviews, which evolved into additional rounds of coding to refine theoretical categories (Maxwell, 2013) and explicate the process of mass reporting as a form of networked harassment on social media.

FINDINGS: WEAPONIZED PLATFORM GOVERNANCE

This unfortunate paradox has been well-documented: decisions regarding the *visibility* of social media creators rest in remarkably *opaque* computation, known to platform companies but not the content producers deeply impacted by algorithmic recommendations (Bishop, 2019; Cotter, 2021; Duffy et al., 2021; Duffy & Meisner, 2023). The creators in this study recounted their experiences with mass reporting as a form of harassment that threatened their presence across various social media platforms, as well as their financial livelihoods. I argue that these creators' experiences with mass reporting represent a shared account of a *weaponization* of platform governance, which I define as the appropriation of infrastructures for platform governance to harass or threaten users, typically with the goal of reducing content visibility or rendering content invisible altogether. In what follows, I unpack the process of mass reporting as it is experienced by social media creators whose relative (in)visibility across platforms has material consequences in their career. More broadly, I use the case of mass reporting to depict a wider trend of tools for platform governance becoming weaponized to target, in many cases, already marginalized creators.

Manipulating reporting infrastructures

User-driven flagging is a common form of content moderation in online communities, but the creators interviewed in this study shared a collective understanding that reporting infrastructures are consistently manipulated and abused for strategic ends, including harassment, reductions in content visibility, and account suspensions. Rather than using a platform's community guidelines as a benchmark for content reports, mass reporting relies on a bad-faith use of reporting infrastructures solely for disruption and harassment. Jacob, a gay creator whose content was mass reported by political conservatives, recounted his initial concern:

> It happened the day I went viral, where I hit on an issue that was currently being promoted by Donald Trump. So, I got a lot of Trump followers following me, and then they'd look at what they're following, and I just got a ton of things flagged that day. I was like, “Oh, no. Is going viral going to take me down?”

Like Jacob, several interviewees had content removed or accounts suspended after engaging with particular groups or networks online, typically those that were not familiar with the creator. Incidents such as these begin with a single post sparking a series of content violation reports, which often also attracts reports on a substantial amount of a creator's previous content, presumably in an effort on the part of harassers to have the account—and not just a single post—suspended or banned. After making comments about the musical group One Direction on TikTok, Allison, an aspiring comedian, experienced a flurry of reports layered with threats on her life. She shared, “Once you hit a nerve maybe with a certain fandom or a certain group of people and they mass report you, then you're at the mercy of any group that decides that they disagree with your video.”

Given that many creators experience myriad forms of harassment, it is unsurprising that some creators compared their own experiences reporting others' inappropriate content with their experiences *receiving* faulty reports on their own content. Antonia explained this double standard she experiences on Instagram:

> All the people that I've reported were sending me pictures of their penis or masturbation videos from strangers and all these sugar daddies that keep trying to scam me, which has been going on for two years. Every time I report them, the review comes back and says, “We've reviewed this account and they're not violating community standards.” Okay, so you can send masturbation videos to strangers, that's okay for your platform, but if I want to post an artistic nude, my account's going to be deleted. Yes, obviously we're following the same guidebook here.

On TikTok, Katrina echoed perceptions of inconsistency and unfairness. She said, “[TikTok's] reporting system allows for mass reporting, and people know it. But what I have noticed is that there are certain things that never get taken down.” These experiences signal that mass reporting is felt as an unfair punishment against creators that also produces feelings of distrust and weakens confidence in platform governance in general.

In an effort to mitigate the potentially damaging consequences of a mass reporting campaign, many creators attempt to mobilize their audiences to help preserve a creator's presence and public image on a platform amid ongoing harassment. This resulting back-and-forth between a creator's advocates and enemies—or, at least, those who spearhead mass reporting campaigns—results in strained relations between platform companies and creators who feel unsupported when facing seemingly endless harassment. Moreover, this process adds to an already flooded pipeline of content violation reports across multiple platforms in the most severe cases. Aashna, a creator who experienced mass reporting on both TikTok and Instagram, shared:

> My whole community had to do damage control for me for five days before I felt safe enough to come back online. I reported these [harassment videos] to TikTok, and you would think they would get taken down. A lot of them did get taken down … and some of them they just wouldn't take down. I'm talking 100+ people in my community reporting these videos for slander and bullying and harassment, and TikTok would not take them down. I blocked them so they couldn't access my content anymore, reported all their stuff, and then they found me on Instagram and started harassing me on Instagram.

Aashna's story captures both dimensions of the current problem with reporting infrastructures on social media platforms: they are so fraught with abuse and manipulation that they have become ineffective, and simultaneously, because traditional content reporting (i.e., nonautomated reporting) avenues are perceived as futile, more abuse and system gaming will carry on undetected in platform communities.

Removing posts and punishing creators

In addition to rendering reporting infrastructures ineffective, mass reporting punishes creators, forcing them out of platform communities and damaging the ability of identity-based and marginalized communities to foster safe spaces for social support. The creators in this study experienced similar patterns of confusion and frustration when attempting to understand platforms' punitive measures based on—what these creators consider to be—baseless content violation reports. Charlie, who was mass reported on TikTok on several occasions, said, “Maybe once or twice, I was able to get it reviewed, but most of the time, I would just be summarily banned. It was with no recourse.” For other creators in this study, the punishment and appeal process resembled an enduring rollercoaster. Athena explained:

> Clearly, someone is complaining. I feel like I've been bullied. Someone was complaining about me, complaining about my videos, and eventually enough people complained. Then, your videos get removed, and your account gets banned and taken down. So, I appealed. You might get an apology and told your video will be put back up, but that still counts for a deleted video. So, the next video that gets [reported], you're banned, because you're showing up with three deleted videos.

Thus, even when appeals are successful and content moderation decisions are reversed, future moderation decisions are premised on all prior incidents. Perpetrators of mass reporting campaigns can therefore trust that even false reports with successful appeals can negatively impact a creator being targeted by the reporting effort. After a similar series of events, TikToker Julia shared that she received constant reports on her content for weeks because successful appeals only resolve past incidents and cannot prevent future incidents of mass reporting. Yet, despite the frustrating series of escalating punishments from platform companies, other creators in this study had less success with appeals and communicating with support teams. To make matters worse for suspended or banned creators, Molly, a cosplayer on TikTok, clarified the financial stakes: “Not only do they take your content, but they scam people into joining the Creator Fund where you get paid for every view if you reach [a threshold], and then if your account gets banned, they keep your money.”

Given that marginalized content creators are already vulnerable in online communities, it is unsurprising that highly visible creators—or those striving to be visible—face harassment on a daily basis. Although all creators are negatively impacted by mass reporting, marginalized creators must consider other vulnerabilities when delicately navigating mass reporting and threats to their community's presence on social media. Hanna, a sex worker who is careful to post content strictly in line with TikTok community guidelines, said, “The systems cannot divorce being from a marginalized identity or sexuality from fetishization and sex in the carnal sense.” For other marginalized users, mass reporting campaigns produced worry for their communities. Nadia shared:

> People think, “Oh, it's just an app,” but that's all we have as fat people. That is all we have. And so that's why I'm so tired of talking about this subject matter. I'm so tired of fighting with Instagram. I gave up on Facebook. I totally gave that up. But with Instagram, I'm fighting because this is what is at stake, you know?

Taken together, social media creators from an array of backgrounds, content genres, and platforms recounted patterned experiences with mass reporting that resulted in individual targeting, harassment, and punishment. Because these harassment campaigns are often coordinated on third-party websites, they also follow creators across platforms, aiming to diminish their visibility as creators—as professional creative workers on multiple platforms—not just as an individual account on a single platform.

Reifying platform opacity and precarity

For years, creators have criticized platforms for their overreliance on opaque algorithmic systems that govern conduct, reward and punish content, and serve as a foundational organizing mechanism for many of today's most popular social media platforms. Those orchestrating mass reporting campaigns prey on creators who (often) are already struggling to produce content that will succeed by algorithmic standards, and mass reporting only amplifies confusion, frustration, and precarity for social media creators. In addition to wider critiques of transparency in the content moderation and appeals processes, the motivations behind coordinated flagging efforts are not always clear. Krista speculated about the ways her content is flagged, though she remains unsure despite numerous content takedowns:

> It's really hard to tell, and they don't tell you why something is taken down specifically. And [removal] happens anywhere from five minutes after posting a video, to five weeks after posting a video. So, I would imagine the posts that come down quickly are probably bots scanning, some kind of computer-generated censorship, whereas the posts that come down weeks later, I would imagine are probably human reports. That's what I suspect, but I really have no idea.

In Krista's case, mass reporting confirmed her perceptions of opacity and confusion surrounding how reports are identified, issued, and upheld. For many creators like Krista, their experiences with mass reporting only reinforced their lack of faith in platform companies in carrying out their governance responsibilities. This lack of transparency is coupled with an alarming lack of communication in the face of a coordinated harassment effort. Claire explained how creators are left in the dark while waiting to know the outcomes of an appeal review:

> No one communicates. We all hit TikTok going, “What's going on?,” and then we all get the bot responses of “Cindy from TikTok,” and “Greg from TikTok,” but it's a form letter because we all get the exact same response.

In the fallout after a creator experiences mass reporting, they are met with an arduous, confusing process to regain full access to their content and accounts. In some cases, interviewees had *still* not regained access by the time of our conversation. During this indefinite waiting period, creators receive empty communication—if any at all—from the platform company and no ongoing support as they navigate the status of their career without access to the platforms hosting their content. Although most creators in this study post content across numerous social media platforms, a lapse in visibility for weeks on a single platform can pose material challenges to creators seeking to expand their reach with each passing day, week, and month.

Creators, of course, are not a monolithic group. They may use their content creation and social media presence as a full-time career or as a supplement to an offline career. Tatum, a small business owner who markets her products through TikTok, experienced mass reporting and described how her content takedowns had material consequences for her and her employees:

> So, the stakes are high. My employees have taken these jobs with me under the assumption that I'm going to be able to provide paychecks for them. And if I take too many risks with our social media accounts and get taken down or get our accounts deleted, I'm not able to recover them. That's literally money running out the door, and that's not money for my pocket that I'm talking about. It's my employees' paychecks, their livelihoods, their health insurance, their paid time off, their vacation.

Given the convergence of creator cultures and practices across other, seemingly unrelated industries, Tatum's experience as a small business owner deeply impacted by sociotechnical harassment, such as mass reporting, will likely become all the more common in the years ahead. Even without frequent waves of mass reporting, social media creators are an underresourced, undersupported class of workers whose work was already fraught with precarity and opacity. Mass reporting only reifies concerns about these problems for creators and, for many, indefinitely leaves them without any answers from platform companies whose governance infrastructures have been deployed against them.

DISCUSSION

This study empirically explored mass reporting on social media as a form of coordinated harassment against social media creators. My interviews with creators who have experienced mass reporting demonstrated how the infrastructures of platform governance have become *weaponized* against creators to facilitate content takedowns and render their content invisible. Moreover, mass reporting simultaneously damaged the trust—albeit already dwindling—creators place in platform governance, particularly as would be expected during cases of harassment. I documented three dimensions of mass reporting as they unfolded for creators: manipulating reporting infrastructures, removing posts and punishing creators, and reifying platform opacity and precarity. This study unpacks mass reporting as a sociotechnical process that introduces vulnerabilities for both social media creators and platforms' abilities to govern their networked communities more broadly.

This analysis offers meaningful implications for the study of platform governance in the creator economy. Previous research in this area focuses largely on the role of algorithmic recommendations in disciplining the professional activities and labor of social media creators. Yet, reporting and flagging systems are essential to content moderation apparatuses at large social media platforms (Crawford & Gillespie, 2016; Gillespie, 2018), though they receive comparatively less attention in recent studies of content moderation. Amid the substantial attention directed at the mediating role of algorithms on digital platforms, contemporary problems like mass reporting compel deeper consideration of the enduring role of the *human* in complex technical challenges. This study broadens our understanding of the types of sociotechnical challenges creators face as they strive to become eminently visible, particularly for creators who are routinely marginalized and harassed based on their identities.

I build from the case of mass reporting to argue that the infrastructures for platform governance on sites like Instagram, TikTok, Twitch, and YouTube have become increasingly *weaponized* against social media creators, such that the avenues for protecting creators from harassment have become the very channels of harassment harming creators. These waves of automated flagging also disrupt the few channels that creators retain for communicating with platform support operations, much of which is automated or handled through—per these creators' accounts—unanswered e-mails.

The community guidelines on Instagram, TikTok, Twitch, and YouTube depict rules and resources for maintaining "positive" discourse on their platforms but struggle to articulate how these regulatory systems become easily gamed and manipulated by groups on (or off) their platforms. Moreover, by emphasizing the supposedly democratic infrastructures for reporting policy-violating content, platform companies fail to acknowledge patterns and hierarchies within accounts being continually reported and accounts filing the copious reports.

This study also sheds light on the ways in which mass reporting intensifies power asymmetries between platforms and creators in the social media economy. Although creators frequently express frustration with the information concealed by platform companies (see Cotter, 2021), creators—much like many other workers across industries—have a baseline desire to work in digital environments that promote safety and clear, enforced policies for community governance. Mass reporting, in its attempt to harm the visibility of creators, also harms the very systems that are used to govern online communities. By damaging the legitimacy of abuse reporting systems, all platform users, not just social media creators, are less safe in their online communities. But while the emotional burdens may be similar for many platform users, creators must manage the added stakes of economic precarity on top of the emotional tolls of harassment. Adding insult to injury, the inequities of creator harassment appear to operate within and through the inequities and patterns of marginalization in society at large. Future research should next take up the question of organized resistance by marginalized groups within the creator economy.

Additionally, this study found that social media creators spend weeks and months at a time waiting on responses from platform companies after content takedowns and account bans, yet studies have only documented the frustration creators feel *against* appeals processes and not their experiences pursuing visibility in resistance to the protracted waiting process in which platform companies have the upper hand. Although other studies of challenges to platform governance have resulted in the widespread reckoning of a site's purpose and values (e.g., Tumblr's infamous #NSFW content ban; Sybert, 2022), mass reporting goes beyond individual content removals or censorship by deplatforming accounts altogether, often temporarily but in some cases indefinitely. This poses practical challenges for creative workers who have lost access to at least one hosting platform for their labor and are attempting to negotiate their return through a broken, fragile, and easily weaponized governance system. Researchers should study these spaces of time amid deplatforming as spaces of coordination and resistance for social media creators, as occasionally alluded to by the creators interviewed in this study. By examining the strategies creators adopt during opaque and lengthy appeals processes at platform companies, we can improve the understanding of creators' resistance practices and coordination as a class of workers. Future research in this area should also examine other instances of manipulations of platform governance that seek to shut down or backlog content moderation and community management processes.

In closing, this study documented the process of mass reporting on social media from the experiences of social media creators who had been victimized by repeated incidents of coordinated flagging firsthand. These findings reveal a new dimension of precarity facing social media creators when the infrastructures they rely on for platform governance are weaponized against them and, in that process, rendered ineffective for helping them regain a presence on the platform. At a time when social media creators are struggling to achieve visibility in a competitive and complicated algorithmic media environment, this study adds an additional layer of challenges for creators, perhaps most notably those whose mere identity or content genre sparks antagonism by groups of hostile networked actors.

ACKNOWLEDGMENTS

Participant compensation and transcription support for this research were funded by the Cornell Center for the Study of Inequality at Cornell University. The author wishes to thank all research participants for sharing their time and expertise; Brooke Erin Duffy for her feedback and guidance; Emily Wolfman and Kate Maxon for their superb research assistance; the New Media & Society Working Group at Cornell University; special issue editors James Fitzgerald and Ayse Lokmanoglu and the anonymous reviewers; and the fellow authors in the special issue for their generous feedback on this work.

ORCID

Colten Meisner http://orcid.org/0000-0001-7398-2269

REFERENCES

Abidin, C. (2016). Visibility labour: Engaging with influencers' fashion brands and #OOTD advertorial campaigns on Instagram. *Media International Australia*, *161*(1), 86–100. https://doi.org/10.1177/1329878X16665177

Are, C. (2022). The shadowban cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, *22*(8), 2002–2019. https://doi.org/10.1080/14680777.2021.1928259

Are, C. (2023). An autoethnography of automated powerlessness: Lacking platform affordances in Instagram and TikTok account deletions. *Media, Culture & Society*, *45*(4), 822–840. https://doi.org/10.1177/01634437221140531

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In G. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Proceedings of the International Conference on Social Informatics*. https://doi.org/10.1007/978-3-319-67256-4_32

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. *New Media & Society*, *21*(11–12), 2589–2606. https://doi.org/10.1177/1461444819854731

Clark-Flory, T. (2019, April 17). *A troll's alleged attempt to purge porn performers from Instagram*. Jezebel. https://jezebel.com/a-trolls-alleged-attempt-to-purge-porn-performers-from-1833940198

Contreras, B. (2021, December 3). TikTok creators say they lose videos through mass reporting. *The Los Angeles Times*. https://www.latimes.com/business/technology/story/2021-12-03/inside-tiktoks-mass-reporting-problem

Cotter, K. (2021). "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, *26*(6), 1226–1243. https://doi.org/10.1080/1369118X.2021.1994624

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Duffy, B. E., & Meisner, C. (2023). Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society*, *45*(2), 285–304. https://doi.org/10.1177/016344372211119

Duffy, B. E., Pinch, A., Sannon, S., & Sawey, M. (2021). The nested precarities of creative labor on social media. *Social Media + Society*, *7*(2), 1–12. https://doi.org/10.1177/20563051211021368

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

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, *8*(3), 1–13. https://doi.org/10.1177/20563051221117552

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). 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*(CSCW2), 1–35. https://doi.org/10.1145/3479610

Kayser-Bril, N. (2021, February 1). The insta-mafia: How crooks mass-report users for profit. *AlgorithmWatch*. https://algorithmwatch.org/en/facebook-instagram-mass-report/

Kou, Y., & Gui, X. (2021). Flag and flaggability in automated moderation: The case of reporting toxic behavior in an online game community. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445279

Marwick, A. E. (2021). Morally motivated networked harassment as normative reinforcement. *Social Media + Society*, *7*(2), 1–13. https://doi.org/10.1177/20563051211021378

Marwick, A. E., & Caplan, R. (2018). Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies*, *18*(4), 543–559. https://doi.org/10.1080/14680777.2018.1450568

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346. https://doi.org/10.1177/1461444815608807

Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media + Society*, *5*(2), 205630511983677. https://doi.org/10.1177/2056305119836778

Maxwell, J. A. (2013). *Qualitative research design: An interactive approach* (3rd ed.). Sage.

Napoli, P., & Caplan, R. (2017). Why media companies insist they're not media companies, why they're wrong, and why it matters. *First Monday*, *22*(5). https://doi.org/10.5210/fm.v22i5.7051

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Sybert, J. (2022). The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *New Media & Society*, *24*(10), 2311–2331. https://doi.org/10.1177/1461444821996715

Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O. (2022). (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society*, 1–22. https://doi.org/10.1177/14614448221109804

Thomas, K., Kelley, P. G., Consolvo, S., Samermit, P., & Bursztein, E. (2022). “It's common and a part of being a content creator”: Understanding how creators experience and cope with hate and harassment online. In *CHI conference on human factors in computing systems* (pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3491102.3501879

Uttarapong, J., Cai, J., & Wohn, D. Y. (2021). Harassment experiences of women and LGBTQ live streamers and how they handled negativity. In *ACM international conference on interactive media experiences* (pp. 7–19). Association for Computing Machinery. https://doi.org/10.1145/3452918.3458794

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

White, M. (2012). *Buy it now: Lessons from eBay*. Duke University Press.

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287

How to cite this article: Meisner, C. (2023). The weaponization of platform governance: Mass reporting and algorithmic punishments in the creator economy. *Policy & Internet*, 1–12. https://doi.org/10.1002/poi3.359