Submitted version of: **Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. Proc. ACM Hum.-Comput. Interact. 4, CSCW1, Article 040 (May 2020), 28 pages.**
DOI: https://doi.org/10.1145/339284

Conformity of Mental Illness through Content Moderation

For individuals with mental illness, social media are considered spaces for sharing and connection. However, not all expressions of mental illness are treated equally on these platforms. Different aggregates of human and technical control are used to report and ban content, accounts, and communities. Through two years of digital ethnography, including interviews with people with eating disorders, we examine the experience of content moderation. Our analysis shows that the practices of moderation available on different platforms have particular consequences for members of marginalized groups, who are pressured to conform *and* compelled to resist, such as through the subversion of platform features and collective action. Above all, we argue that platform moderation is enmeshed with wider processes of conformity to mental illness and body image. Practices of moderation reassert certain bodies and experiences as 'normal' and valued, while rejecting others. At the same time, practices of navigation around and resistance to these normative pressures further inscribe individuals as on the margins. We discuss changes to the ways that platforms handle content related to eating disorders by drawing on the concept of multiplicity to inform design.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Eating disorders, content moderation, mental illness, social media, pro-eating disorder, pro-ED.

ACM Reference Format:
. 2020. Conformity of Mental Illness through Content Moderation. *J. ACM* 37, 4, Article 111 (August 2020), 23 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Participation of diverse groups of people on social media platforms such as Facebook, Instagram, and Reddit, occupies a large contingent of work in Computer-Supported Cooperative Work (CSCW). Research addresses the proliferation of networks and communities across these platforms, as well as the content of discussions and practices of sharing among members [1, 4, 34, 58, 68, 92]. Emergent within this literature is an emphasis on understanding the practice of content moderation and associated experiences. As Gillespie argues, content moderation is central to what online platforms do [44]. Moderation of participation and discussion has been studied within general contexts, such as Reddit [59, 61], as well as specific ones, including examination of hate speech and online harassment [19, 99, 115, 116]. Much discussion in this domain involves identifying specific topics of conversations [21, 116], determining which topics are right or wrong to encourage [24, 40, 46, 111], and where the line should be drawn between manual and automated forms of regulation [60, 112]. In the CSCW- and Human-Computer Interaction (HCI)-related literature, as well as publicity from large tech firms [100], the moderation of individuals and groups has been treated as a necessity for the greater good.

In this paper, we aim to approach the topic of content moderation from another perspective. We closely attend to how moderation happens and what the consequences of moderation are

Author's address:

Submitted 16 Oct 2019. Published May 2020

for members of marginalized groups expressing non-dominant narratives. We argue that the mechanisms of moderation afforded on social media platforms exert an active force, producing and reproducing a conformity to particular norms and values. This is not necessarily a conformity to the rules formally documented in a platform's standards and guidelines [35, 44, 96], but, rather, refers to an emergent and tacit set of norms and values that get negotiated and enforced by a platform's members. Our interest, then, is in how the distinctive capacities of social media platforms make a particular kind of moderation possible. As this is a moderation that precipitates and reinforces an emerging but still narrow set of norms and values, it raises larger questions of how moderation—as performed on social media platforms—shapes online participation and delineates whose voices are permitted online and whose are not.

Here, we examine the social and technical practices of content moderation on social media platforms as they relate to individuals with eating disorders. The work that follows is grounded in two years of digital ethnography, most recently focusing on the experiences of individuals with eating disorders across an ecosystem of social media platforms. In addition to analyzing online content, we interviewed 20 individuals with eating disorders who reported having content removed from social media platforms, including Facebook, Instagram, Reddit, Tumblr, and Twitter. We show through this study that the pressures of moderation can have damaging consequences, especially for marginalized groups. These consequences include loss, labor, and displacement, as well as wider processes that reinforce ideas around which versions of mental illness and body image are sanctioned as 'normal' and 'acceptable' in online spaces. As Dani, one of our participants, said, *"We totally shouldn't mute people out of mental illness communities because we don't like what they post."* There's harm in having *"your thoughts and your emotions just shut down because it doesn't align with what, I guess, people who are, I air quote, 'normal' want to see."* We will show that resistances arise in response to these many consequences and to the effects of being marginalized. Individuals and groups find ways to subvert oppressive platform processes through, for example, the creation of different user accounts or establishment of splinter communities forged through their own ingroup, grassroots processes of community moderation.

We make three primary contributions. The first is a detailed account of how members of a marginalized group—individuals with eating disorders—experience content moderation, extending prior work in this space [18, 59, 61, 89]. Although content moderation is typically conceptualized as necessary for the greater good of online communities (e.g., preventing harassment), its potential harms are not well-understood or documented. Our analysis reveals the ways in which content moderation has consequences, sometimes severe, for people with eating disorders, including loss of personal content (e.g., used for self-reflection) and community support, as well as creating additional restorative work for people who have been subject to moderation. Second, we turn to the notion of conformity as a way of understanding the broader social and technical practices of content moderation. We discuss how content moderation contributes to wider processes of conformity, which, in this context, construct particular versions of mental illness and body image as legitimate while rejecting and silencing others. Third, as a counterpoint to conformity, we reflect on what it means to design for multiplicity in online social platforms and articulate directions for future work aimed at creating more equitable online spaces.

2 RELATED WORK

Our work builds on a growing body of literature related to social media and content moderation, how people with eating disorders engage online, and the ways in which members of marginalized groups participate and interact on social media and other online spaces.

Submitted 16 Oct 2019. Published May 2020

2.1 Content Moderation on Social Media

A large body of work within CSCW- and HCI-related literature examines content moderation in the context of social media and online communities [7, 19, 20, 44, 59–63, 71, 84, 89, 111, 112]. Practices of moderation aim to facilitate quality content, civil discussion, and, generally speaking, online spaces where individuals can engage and participate without overt fear of abuse, harassment, or accidental viewing of violent or illegal activity [71, 72]. Throughout this paper, when we refer to content moderation, we refer to "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [48]. What we call platform moderation, others have termed commercial content moderation [105]. This practice of moderation involves the organized ways in which content produced by social media users is subject to surveillance, report, review, and removal [89]. These practices often rely on decisions passed down by dispersed groups of outsourced laborers [47, 105].

Though mechanisms behind content moderation are largely proprietary and private (i.e., a black box [60]), some researchers have illuminated the underpinnings of these sociotechnical processes [44, 47, 89, 105]. Broadly, content moderation may involve automated systems, community flagging and reporting [23], and outsourced labor [47, 104, 105]. Several social media platforms, including Reddit and Facebook (e.g., subreddits, Facebook groups), also rely on community moderators—at times, with automated systems—to manage groups of individuals with similar interests, as well as transient visitors [60, 63, 66, 84, 112]. We distinguish this instance of moderation, in which moderators and other members of communities engage in shaping (i.e., moderating) particular forms of participation online, from platform moderation. However, as we argue, practices of platform and ingroup community moderation are entangled.

Given the pervasiveness of content moderation, a growing area of interest involves understanding the experience of being moderated [59, 61, 62, 89]. This research thread speaks to the frustration and, at times, confusion of having content removed. Though marginalized communities and groups of people are not highlighted currently in this body of work, researchers have suggested that content moderation may have more detrimental effects on their members [62, 89]. The present study helps bridge this gap in the literature by engaging with a particular marginalized group (i.e., individuals with eating disorders) through digital ethnography, including online observation and interviews. In addition to demonstrating the harms of content moderation in this context, we animate its role in constituting eating disorders and, as we detail in the discussion, illness narratives online.

2.2 Content Moderation and Eating Disorders Online

Researchers have also studied content moderation as it relates to eating disorders. This work typically engages with deviant (i.e., rule-breaking) content from pro-eating disorder (pro-ED) communities. Research in this domain has used machine learning techniques to characterize types of content removed [15, 16], as well as behavioral responses to moderation, including the ways that individuals use platform features to circumvent banned content, such as hashtags [18, 43]. Findings from these works, in part, provide valuable insight as to how platform practices of moderation (i.e., particularly the banning of hashtags) amplify existing challenges to moderation and may inadvertently overlook others. For example, Chancellor and colleagues [18] found that attempts to moderate certain types of eating disorder content through hashtag bans resulted in a broader diversity and lexical variation of hashtags—thus adding to challenges surrounding moderation via hashtags. Gerrard [43], similarly, detailed additional limitations to practices of hashtag-based moderation, including the ways in which recommender system actively circulates pro-ED content. Due to these pitfalls of platform moderation, researchers note that alternatives are necessary

[18, 33, 43]. In this paper, we extend these prior works through an empirical study of the experience of content moderation and a subsequent discussion detailing new avenues for design

Content moderation is not, of course, the lone interest for researchers examining eating disorders online. Prior works detail a large and diverse spectrum of inquiries, including characterizations of content [8, 14, 25, 45, 64, 95, 97, 117], information-seeking behaviors [9, 36, 78, 90], recovery likelihood [17], and ethical concerns, including those with respect to censorship [113]. With respect to research focusing on pro-ED content and communities, recommendations may include novel forms of moderation [15, 25], such as automated systems to assist human moderators, and health interventions [16]. Though recovery and pro-recovery communities [25, 70] are included in this domain, the majority of research examines pro-ED content and communities online. Here, we take a broader view by including a diversity of experiences. Additionally, many prior studies do not engage directly (i.e., through interviews) with the communities they observe and plan to serve. First-person accounts are vital to better understand the complexity of eating disorders online. With this paper, we build on these earlier studies with interviews and attention to the adverse effects of content moderation on individuals with eating disorders.

2.3 "On the Margins" of Social Media and Online Communities

The contemporary experience of living with an eating disorder cannot be understood without considering the historical context of mental illness. Historically, individuals living with mental illness have encountered stigma, social ostracization, and forms of oppression, including forced institutionalization [37, 75, 110]. Specific to eating disorders, research has found that anorexia and bulimia are significantly more stigmatized than depression [106], and that eating disorders are associated with a variety of stigma and negative stereotypes dependent on the specific diagnosis (e.g., anorexia, bulimia, and binge eating disorder) [102]. In this paper, we join with other mental health and social media researchers in considering the experiences of people with eating disorders—and mental illness, more generally—through a history of marginalization [33, 91, 95]. Situating the experiences of these individuals in the context of marginalization helps us to better attend to power dynamics and differentials, acknowledge labor practices, and contribute to a growing body of literature that examines the marginalization of groups and design for more equitable online experiences [7, 50, 57, 103, 109].

Beyond research examining eating disorders, there is a large body of work that examines how individuals with 'non-normative' identities or behaviors engage and participate online [27, 31, 49, 81, 85]. Though there are benefits to online participation for members of marginalized groups, there are also an array of harms. For example, women, people of color, members of the LGBTQ community, and individuals with mental illness all encounter disproportionate and targeted forms of harassment online [30, 33, 76]. Ongoing research aims to address problems with harassment, such as through work with social organizations, communities, and platforms, including Hollaback [28] and HeartMob [7]. Social media platforms are also invested in understanding and solving problems related to online harassment [73]. However, as Gillespie describes, platform efforts related to reporting and mitigating harassment can themselves contribute to the problem (e.g., such as when individuals organize to use reporting features to flag or report a specific user—or group of users—who they do not agree with or like). Here, we consider how features designed for good (i.e., moderation to support positive user experiences) can work to exclude individuals with eating disorders and contribute to the oppression of a marginalized group online.

3 METHOD

To understand the experience of content moderation for individuals with eating disorders, we conducted a two-year digital ethnography of mental illness across a variety of social media platforms

and online communities. We couple this in-depth analysis of online content with 20 semi-structured interviews with individuals who have or have had eating disorders and experienced the effects of content moderation online.

3.1 Digital Ethnography

We began our digital ethnography in November 2017. At that time, Instagram was our only online site. Initially, our inquiry was focused on understanding multimodal expressions of mental health and illness [77]. We built an initial corpus of posts using hashtags that had been verified in previous research [2, 16, 18] and by manually visiting public posts and accounts. This initial corpus comprised of 2,102 posts. Using this initial corpus as a starting point, we collected a total of 6,223 Instagram posts by tracing through accounts of individuals who had posted, commented, and liked posts in our corpus. This traversal helped by including accounts, posts without hashtags and, indeed, posts without any text at all. Within this corpus, we noticed a number of instances in which individuals who had content removed on Instagram, including content related to self-starvation and self-harm (e.g., cutting), posted about their frustrations with Instagram as a platform.

In addition to Instagram, we also conducted data collection on Reddit. We did this to expand our corpus beyond a single social media platform. Broadening data collection provided a more holistic view of the ecosystem [12, 27] in which people with eating disorders interact, and how this ecosystem changes and is disrupted by platform moderation of content and communities. During November 2018, Reddit issued a series of bans to communities such as r/ProED, r/ProEDMemes, and r/ProEDAdults. Following this incident, we observed how a number of banned subreddit members joined other social media and online communities. At this time, we also collected public posts discussing the platform's decision to ban these subreddits on Reddit and other relevant online spaces, including online communities, individual blogs, and social media. In this paper we refer to banned eating disorder support communities on Reddit by their names. We do this as a form of activism to raise awareness about an unjust practice—namely, terminating a community that provided support for a marginalized group. Our digital ethnography involves currently active and quarantined subreddits, as well as several smaller, online communities that are not housed on social media platforms. To preserve the privacy of these communities and their members, we do not name them.

We also collected posts from Tumblr beginning in June 2019. We began our manual crawl through Tumblr using eating disorder hashtags, such as those verified in previous research and occurring alongside these hashtags on posts and within accounts. In total, we collected an additional 368 threads and posts from Reddit and Tumblr. These data were used to inform our line of questioning for interviews and supplement analysis. When presenting these data, we alter the wording of posts so that they are not easily searchable or identifiable.

3.2 Interviews

We conducted semi-structured interviews with 20 adults (ages 18 – 57; M=29) with eating disorders who had content related to their disorders removed from online communities and social media platforms. Though eating disorders can impact anyone [97], regardless of any particular facet of identity, only three participants in our study identified as male (17 female). This is not to suggest that eating disorders are more prevalent or significant for women, only that our methods of recruiting did not adequately reach out to or engage with other individuals. With respect to race, eating disorders often run the risk of being associated predominantly with white women [69]. This is not the reality. While the majority of our participants were white (n=12), six were African-American, one was Hispanic, and one identified as multi-ethnic.

Eligibility for this study was not contingent on a diagnosis. However, barring diagnosis, participants were required to identify as having an eating disorder. We invited individuals living with and in recovery from eating disorders to participate. As such, we have a broad spread of experiences represented by our participants. For example, several of our participants described being in recovery, while others were relapsing at the time of the interview or had grown accustomed to living with their disorder. We interviewed individuals who were members of pro-ED communities, as well as individuals who were members of pro-recovery or diet communities. The thread connecting our participants were their experiences, even those in the past, with content moderation. The content removal experienced by our participants included posts, accounts, and communities.

We recruited participants from an online eating disorder support community, Reddit, Craigslist, and snowball sampling. We issued a pre-interview phone screener, where we called participants to verify their age, eating disorder status, and experience with content moderation. Interviews lasted an average of 45 minutes and were held over the phone (n=18) or in-person (n=2). During the interview, we discussed topics related to experiences with online eating disorder accounts and communities, content removal, reactions to content removal, support resources, and opportunities for platform redesign. Interviews were audio recorded and transcribed for data analysis. Participants received a $30 Amazon gift card or $30 in cash. Additionally, mindful of the health and wellbeing of our participants, we included a mental health consultant as a member of our study team. This individual did not participate in data analysis, but was available for participant outreach. When referencing our participants throughout the paper, we use pseudonyms.

3.3 Data Analysis

Our approach to data analysis follows a constructivist grounded theory process, where members of the research team developed themes through iterative coding, memo writing, and constant comparison of data to developed concepts [22]. We began analysis with our early interviews and used initial findings to shape the trajectory of subsequent interviews and recruitment. We developed themes by analyzing interviews in tandem with the online content we had collected. Preliminary themes included types and motivations for posting content that was eventually moderated, receipt of news (i.e., how participants came to know their content had been moderated), sensemaking around moderation, consequences of moderation, subversive practices and resistance, and tensions with coexistence (i.e., how individuals navigate eating disorder communities that may include trigger content). Through our analysis, we began to understand the ways in which harm can be caused by good intentions (e.g., content moderation and support resources) and the ways that individuals push back on oppressive practices, as well as how they participate online in such a way to support the diversity of experiences with eating disorders.

As part of our method, we are mindful of how the analytic frame of marginalization requires accounting for and reflecting on how our expectations, values, and norms as researchers, and as individuals within society, differ from those of our participants and online posters [11, 52]. It is through this analytic framing that we began to see the concept of conformity take central focus in our analysis and understanding of content moderation. We return to this in the discussion.

4 FINDINGS

Through our analysis, we show how content moderation involves the interplay of social norms and technical features of a platform that work to erase individuals and remove support, create new labor by encouraging responses and resistance, and shape community-led practices of moderation. To set the scene for our findings, we first walk through a case with one of our participants that illustrates how content moderation works in this context.

Dani, now 20 years old, has participated on social media and online communities for nearly a decade. Though her personal experience with eating disorders was not the only content she shared online, it did specifically result in account bans on both Tumblr and Instagram. With respect to Tumblr, prior to the termination of her account, Dani used a number of strategies to manage her public eating disorder blog and limit unwarranted attention. She avoided using features that could establish links to other content or aid in platform search and providing tips or advice to other users (i.e., *"telling people you should do this"*). Despite these strategies, Dani felt like she was *"walking on eggshells"* whenever she posted. Her sensitivity to the workings of Tumblr (e.g., its facilities for linking and connecting content) was motivated by wanting to maintain a highly personal blog detailing her own sense of self and body image, while, at the same time, wanting to escape criticism and overly stringent platform moderation. Specifically, she *"didn't want people to come crucify me because I was talking about, you know, the part of eating disorders that nobody wants to see. That nobody wants to hear."*

Despite Dani's strategic use of Tumblr, her eating disorder blog attracted attention. A year into managing this blog, Dani received an "aggressive" anonymous message asking her to delete an unspecified post about body image. *"I didn't know exactly which post they were talking about,"* she said. *"[T]hat wasn't the first time I posted about me not liking the way I looked... So, for a moment I sat and stared at the [message], and I was like, 'What? Which one?"'* Rather than remove any posts, Dani sent a message back to the anonymous user, telling them to "just block" her. Shortly after, Dani's account was terminated by Tumblr. An email from Tumblr's support team notified her the eating disorder blog had been deleted for *"violating their terms"* and, though it invited appeal, Dani's efforts to receive an explanation and reinstate her content remain unanswered. While what eventually triggered the ban is unclear, Dani placed blame with the anonymous user who messaged her earlier in the day. However, it may have been another, or even an automated content reporting system, that was ultimately responsible. Despite being subjected to regulation, Dani subverted the ban on Tumblr by creating a new account and, ultimately, finding new online communities, including those off of social media, to join. However, even with new accounts and online spaces, Dani's experience of being banned shaped her future interactions online, including practices of disclosure. She explained:

> *"I'm not as talkative anymore... I just kind of lurk... I know there's still people posting about eating disorders on there, but, when I see a post from them, I immediately get nervous saying, you know, if I interact with this person...someone is going to find my account and find a reason to make me disappear."*

We see in Dani's case how a range of sociotechnical mechanisms and practices can work together to monitor and regulate content. This not only controls what and where some individuals post, but also shapes what constitutes appropriate or acceptable versions of having an eating disorder online. For Dani, we see that content moderation has serious consequences, including reduced social engagement and online expression. Additionally, we see how moderation and its consequences, including the lingering possibility of further sanction, serve to amplify Dani's sense of being the subject of control and of surveillance.

As Dani's case begins to show, individuals experience a multitude of serious consequences following from moderation, leading many to react against and resist platform moderation. Platform moderation, however, is not simply an external force acting on these individuals. It is an interactive process that shapes how groups of individuals with diverse and varied experiences of eating disorders establish their own community-led forms of moderation as part of engaging and participating within online spaces.

Submitted 16 Oct 2019. Published May 2020

4.1 Experiencing Content Moderation as Loss

Throughout our data, and exemplified in Dani's case, we learned of many unintended consequences to moderation when content related to eating disorders is moderated by platforms, including reduced online engagement and loss of community. Marie, discussing an experience with account termination, addressed how, for her, moderation *"was kind of embarrassing."* She *"felt like I was being told I was wrong. Or getting punished when I hadn't done anything. I felt like I hadn't done anything wrong and I was angry about that, as I felt it was unfair."* The initial anger and confusion associated with moderation, as Marie and others in our dataset described, have been detailed in prior research [59]. These—often strong—emotions are entangled with the ways that individuals learn about and make sense of the experience of moderation, which can be confusing due to the lack of transparency and consistency. Marie's comment, in addition to describing her embarrassment and anger, speaks to recent findings detailed by Jhaver et al [59]. Notably, that many individuals who have been moderated feel that they were done so unfairly. Here, rather than focus on perceptions of fairness or emotional responses to moderation, we attend to the various losses, including personal content for reflection and community support, that moderation entails.

Loss of content is central to the experience of being moderated. Platform moderation often involved unsolicited removal of personal posts and accounts, both of which are maintained by and, ultimately, for the individual. As most of our participants were not in the habit of saving content to multiple locations, their content was lost entirely. By removing or deleting this personal content, platforms effectively erase certain experiences and prevent opportunities for reflection and catharsis. Andrea and Dani both equated aspects of their online content with "diary" entries. While access to online-content-cum-diary-posts is valuable at any point during an individual's experience with an eating disorder [80], Andrea talked about how rereading her earlier posts was beneficial during recovery. She said:

> *"I remember I used to post a lot of intrusive thoughts and then, going through recovery, I started having a lot fewer of those. And then there's a lot of elements where you're like, 'Oh, am I in a really bad place?' And then you go back and look at it and you're like, 'Oh, I'm not having 50 obsessive thoughts today about needing to weigh myself...' I can actively see how it's changed or even like at the time too, seeing how it got worse. That was really helpful to me right when I started recovery..."*

Content removal as a practice of moderation suggests that certain experiences with mental illness are invalid and illegitimate [33]. In these instances, moderation can feel like a loss of personal voice or silencing of experience. While many of our participants shared content related to living with an eating disorder, Grace discussed how posts on her Instagram account centered on *"trying to be healthy"* and *"trying to gain my weight back."* Despite this recovery context, Instagram removed a selfie that Grace shared because she looked too thin. The removal of her post from Instagram left Grace feeling sad, ashamed, and *"unworthy to be seen."* This example demonstrates similarities between various types of eating disorder content, such as recovery imagery and thinspiration, and also speaks to difficulties of classification [32], as well as the ways platforms may inadvertently delegitimatize experiences.

Another form of loss that individuals experience as a result of moderation involves loss of community and social support. When platforms moderate content, they may *"[take] away a support system,"* Christy explained. Loss of community, such as through practices related to account and community bans, can lead to social isolation, particularly for individuals who *"don't have anywhere else to go,"* one former member of the now banned r/ProED wrote. As another former member described, the subreddit ban was *"extremely upsetting. So many people used this [subreddit] for help and support. We can't always find that support offline."* Social isolation due to practices of moderation

can effect health. For example, Dani had a few helpful *"people [on Tumblr] that would tell me, you know, 'You're not alone. I'm here to talk,' and stuff like that."* Following the ban of her Tumblr account, Dani lost these meaningful connections, which caused her to feel *"depressed, 'cause I didn't have anyone to talk to."* In addition to depression, we observed instances in which the experience of moderation led to dangerous offline behaviors, including purging. A former member of r/ProED wrote, *"I was really trying to recover... I don't know what to do now. I really feel like purging everything. This is so stupid."* In attempting to remove content classed as non-normative and harmful, platforms can create a downward stream of negative consequences, including loss of social support that, at times, amplifies illness.

Content removal is not the only practice of moderation that results in loss of community. On Reddit, the practice of quarantining suggests that, while certain subreddits are "not prohibited," they are, nevertheless, not normative or socially sanctioned. Quarantine on Reddit is established in several ways, including warning messages, warning screens, and removal from non-subscription feeds and search. Prior to the quarantine of her subreddit, Morgan described how she revived the community to the point where hundreds of people subscribed every few weeks. Following quarantine, new subscriptions to the subreddit, as well as member engagement, have slowed to a halt. Quarantine, Morgan said, *"severely effects the subscribers"* of a community. *"It also makes you not want to talk, really. It kind of feels like you're under watch. Like, the thing you say, that's going to be the next – that's going to be the thing that makes you get banned."* As this example illustrates, content moderation through quarantining can result in loss of participation and constrained expression due to its surveillant property, which ultimately works to constitute which versions of eating disorders are permitted online.

Loss of community, particularly with respect to the removal of community spaces and content from online platforms, also involves the loss of a shared archive of resources. Andrea discussed how the loss of community resources on r/ProED *"totally sucked, because it was stuff that I would go and read if I was having a hard day. Like, someone had posted what to do if you feel like you're going to binge or what to do if you feel like you can't eat today."* Rather than cultivate community-provided resources, when certain content related to eating disorders is moderated, many social media platforms share support helplines—namely, the National Eating Disorder Association (NEDA) helpline. When Marie was provided the NEDA helpline following the termination of her MyFitnessPal account, an account she'd used for nearly a decade, she felt "insulted." Vehemently, she said:

> *"People will just be, like, 'Here's the NEDA helpline. Hail Mary full of grace. The Lord is with thee.' Because they just don't know what else to do. They don't know what else to say. You just sort of start to feel, like, here's the NEDA helpline. Now please go away... Stop having an eating disorder."*

Helpline resources such as these, while beneficial for some, may feel *"unrealistic and unfair"* to many individuals because of how they push recovery—thereby holding up recovery as an ideal. Provision of these resources in tandem with moderation suggests that platforms are forcibly creating loss—by removing opportunities for reflection, spaces for expression, and online networks for support and connection—and filling that absence with a resource list and phone numbers. The compounded losses experienced by individuals with eating disorders lead many to develop strategies of resistance that aim to circumvent or push back on oppressive platform practices.

4.2 Responding to and Resisting Content Moderation

Given the significant and even traumatizing effects of content moderation, individuals with eating disorders respond in a variety of ways. Here, we address responses to moderation through the lens of resistance. By emphasizing individual and collective action, we acknowledge the labor performed

by individuals with eating disorders. Much of this labor relates to the ways that individuals resist oppressive sociotechnical practices in order to raise awareness about and appeal decisions of moderation, rejoin online platforms by creating new accounts and communities, and engage on platforms by mediating the types of content they decide to share.

Because the removal of content through platform mechanisms often plays out in the background, individuals with eating disorders must work to raise awareness about their experiences. For example, when personal content and accounts are deleted, only the individual who posted the content or who owned the account is notified by email. The constant stream of content via personalized social media feeds, such as on Instagram, Twitter, Facebook, and Reddit, makes it difficult for other users, even those within the same communities or networks, to notice that reblogged or reposted content has been banned or that accounts have been muted or removed. To foreground practices of moderation, raise awareness, and confront other posters, particularly those implicated in practices of moderation, individuals with eating disorders may post about their experiences with moderation online. One Instagram user, for example, captioned a post, *"Why are you reporting me? Why do you want to delete my posts? It makes me feel bad. Seriously, just block me."* Dani similarly addressed her Instagram followers, via a secondary account, when her primary account, a private account, was banned:

> *"I hopped over to my second account and said, 'Hey, guys. Someone reported me' ... I made a post saying, 'Hey, guys. I got my account deleted. I don't know which one of you did it, but gee thanks. That really did - that did me a great favor.' Like, 'Thank you so much because that made my day so much better.' I was furious and I could not for the life of me figure out who it was."*

As we see here, rather than change any personal content or settings, such as post or account privacy, individuals may confront their followers and those who come across their account, holding these other social media posters accountable for outcomes of moderation and requesting they cease and desist. Similarly, many former members of r/ProED posted to other subreddits and online communities about their frustration, anger, upset, and outrage at the ban of their support community. In these examples, we see how individuals use social media, sometimes the very same ones from which they had content removed, as platforms to speak out. This suggests that individuals are highly attuned to how social media can be used for activism [3, 53, 55, 74] and the ways in which other posters contribute to practices of platform moderation (e.g., such as via reporting posts).

Individuals use the technical features they have at hand to raise awareness and respond to what they view as unjust and unfair practices. For example, several participants used platform appeal features provided within email messages detailing content bans. However, many of our participants spoke to the idiosyncratic and opaque nature of appeal. Out of all our participants, only Grace had her Instagram account restored—and on the stipulation that she remove all posts in violation of community guidelines and discontinue her prior posting practices. Other participants described appealing content moderation through forms of collective action, including community-led petitions and surveys. As a key instance, former members of r/ProED created a survey to submit to Reddit admin to *"tell them how most people found the [subreddit] to be really helpful. How it made us feel less alone, like we had people who understood. I hope they listen and unban it. If not, at least we can speak out."* Favorable outcomes to these individual and collection actions were rare, however. Appeals, though offering some mitigation to oppression, are still sanctioned and overseen by platforms. Meaning, appeal processes remain inherent within, rather than a check and balance to, platform moderation.

Submitted 16 Oct 2019. Published May 2020

Of course, platform-sanctioned appeals are not the only way that people return to platforms following account and community termination. Sasha, for example, found it easier to create a new Instagram account, which she did "instantly" after her first one was banned. She was then faced, however, with the task of regaining followers and connecting with individuals from other communities of which she was part. For some, such as those on Tumblr, it is common to create a new account following a ban and request for others to not only follow, but to circulate new blog information via reblogging features. In a post on Tumblr, one person wrote, *"Hey, it's [former blog name]. I got deleted, but I had 800+ followers. Can you share this to help me get them back?"* On most social media platforms, following an account ban, there are few, if any, actual barriers to account creation. Platforms, after all, want new users. As such, it is relatively straightforward for individuals to subvert platform features as a form of resistance and to rejoin spaces from which they have been forcibly removed and displaced. Similarly, it is just as straightforward to join new online spaces and communities. By allowing for these practices of resistance, which, ultimately, platforms do through their technical affordances, platforms offload the labor and burden of content, account, and community recreation to the individuals themselves.

Despite practices of resistance, such as the creation of new accounts or community spaces, platform moderation still exists. This means, for some individuals, continued engagement and participation involves changing content-sharing practices and internalizing the norms that practices of platform moderation aim to establish and enforce. Andrea, for example, described hesitancy around posting on a new eating disorder support subreddit following the ban of r/ProED. *"There was definitely something I wanted to post,"* she said. *"And it was, like, I don't know, I feel like it had specific numbers or I was complaining about not being small enough... And I felt like I couldn't post it and I felt like that was frustrating."* This hesitancy and, ultimately, assimilation of platform standards through self-censorship can have, as we show, a chilling effect on behavior [83, 98]. Sasha, following the ban of an Instagram account, created a new account where she posted content that was *"still in the same arena, just not as intense."* Christy, similarly, described how she stopped posting thinspiration. She explained, *"I just save it or archive it now. So, I don't want to risk, like, getting anything banned. So, I just save, archive stuff that you can find it on, like, various eating disorder websites or on Instagram."* Internalizing platform standards to mitigate risk of moderation conceals experience in a way that is similar to how individuals with eating disorders may hide certain behaviors from friends, family, and social others [114]. Concealing mental illness, eating disorders included, has serious consequences [94]. However, in order to exist and participate within certain online spaces, this is exactly what must be done. Notably, even though individuals assimilate to online norms, they do not necessarily change their offline behaviors. As Marie explained, *"[An account ban] isn't gonna make me stop having an eating disorder."* This sentiment, as well as practices surrounding joining new social media or online communities, speaks to the pervasiveness of content and the tenacity of individuals who, indeed, have a right to exist, express, and connect with others.

4.3 Establishing Norms through Community-Led Moderation

An everyday part of engaging with online eating disorder spaces involves, at the very least, brushing elbows with a variety of individuals and content—which can challenge online participation and result in practices of ingroup community moderation. As Marie described, *"different types of people exist in the same spaces. It's muddled at times."* Among our interview participants, individuals described varied diagnoses (e.g., binge eating disorder, anorexia, bulimia), both clinically provided and self-applied, and relationships with eating disorders, including those related to recovery, relapse, and living with a disorder.

Beyond diagnosis or experience, our participants also engaged across a number of platforms and types of communities, including those described in prior research as pro-ED [25, 45, 95]. Pro-ED has

a long history of negative publicity and association. However, as our data suggest, contemporary usage by community members has reconfigured pro-ED to refer to groups of individuals who support people with eating disorders (i.e., *"in favor of—or pro—people with eating disorders"*), rather than being supportive of the disorder itself. For Marie, pro-ED meant, *"I'm dealing with a disorder and I don't want help right now. And I want a place to vent about that. And it's not so much as being, like, give me tips, give me tricks on how to be skinny... It's more just the support."* However, as Howard and Irani describe, participants may have their own personal and political reasons for participating in interviews [56]. Our understanding of pro-ED is, therefore, not solely grounded in interview data, but also in the types of content that come to be socially sanctioned on social media and online platforms by members of pro-ED communities. For example, content related to the difficulties of having an eating disorder, attempts to recover, and replacing certain harmful practices with less harmful, or safer, ones can exist in the same spaces alongside thinspiration, food diaries (i.e., in the context of an eating disorder and recovering from one), and body checks (e.g., progress pictures of weight loss or gain). Given this diversity of content and experiences, individuals work to establish what is 'normal' and what is not through community-led practices of moderation.

Community-led practices of moderation, in part, develop through individual reflection and action with respect to the ways in which diverse, heterogeneous groups of people—with respect to pro-ED communities and the broader ecosystem of eating disorders online—can coexist. For example, sharing content related to the reality of living with an eating disorder may upset or unintentionally trigger others. This also includes certain experiences related to recovery, where individuals may keep detailed food diaries and share successes related to enumerated weight gain. Individuals are aware of the complexities and tensions between types of content and people. As one Tumblr poster wrote:

> *"I feel bad about running an ED blog. Does anyone else ever feel that way? Like, just kind of guilty. At least a little bit. This blog is for me to vent and cope and meet other people with the same issues. But, like. I'm really nervous that how I express myself is going to mess up some other kid."*

This commentary demonstrates tensions between wanting to engage online (e.g., to vent, cope, and connect) and a deep concern regarding the potential to negatively affect others. This concern influences how people post (i.e., how people self-moderate and self-censor), as well as the ways communities self-govern.

Although our informants were willing to engage in community-led moderation and self-governance, social media platforms, by the very implementation of their features, present a number of challenges to these practices. Marie, for example, explained how *"everyone uses the same, like, 12 [eating disorder] tags on Tumblr for everything. So, everything bleeds together."* While this overlap of hashtags blurs boundaries (e.g., between individuals, content, and the potential for classification), it also presents a number of risks, including the overlap of content in detrimental ways. For example, in one Tumblr post, a user wrote, *"I've recently seen a bunch of recovery tags in non-recovery spaces! Do NOT post recovery tags with thinspo!"* Awareness of the ways in which content bleeds together, as well as its potential risks, is not enough to establish or enforce norms around hashtag use. In part, this is due to colloquial usage of hashtags to broaden audience—and, with it, the potential for followers and likes—as well as the decentralized forms that groups and communities of people with eating disorders take on social media.

Even in these muddled, entangled online spaces, individuals can be mindful of one another. Rose, currently in recovery, talked about how she's able to safely access and participate within one of Twitter's eating disorder communities due to the ways that she, and others, make use of content warnings—labels within posts that are separate from hashtags. According to Rose, on Twitter,

content warnings are *"when someone posts something, for them to actually put up a content warning on top of [the tweet]. So, just to say, like, eating disorder or food or weight or, you know, whatever."* In Rose's community, the use of content warnings are *"kind of an unspoken rule."* Though these warnings are not necessarily standard within or across platforms, their presence in Rose's community enable her and others to *"safely and, in a positive way, access Twitter—is if I have those warnings, so that I can scroll past, if I need to without being triggered to start doing unhealthy behaviors."* Other online communities, such as various subreddits, may also have community practices around flairing or labeling posts. Similarly, many smaller online communities for individuals with eating disorders build community spaces on traditional forums, such as those that allow for category-specific subforums (e.g., Anorexia, Recovery, Thinspiration). In these instances, content has designated spaces. As our participants shared, some online communities are successful at upholding the organization of these spaces, both in part due to an active moderation team, as well as the willingness of members to post in appropriate spaces and call attention to those who do not. Community-led approaches, when successful, *"makes you feel safe,"* Marie mentioned. Ingroup community moderation can facilitate safety and a diversity of content in ways that practices of moderation enforced by platforms do not.

Nevertheless, practices of community-led moderation interact with practices of platform moderation as members work to establish and enforce community norms. Here, we provide the example of harm reduction to illustrate differences between the ways that community-led and platform moderation interact to regulate content. Harm reduction refers to materials or resources that help individuals take care of themselves while living with an eating disorder. For example, harm reduction involves reminding individuals to hydrate during episodes of self-starvation and to not brush their teeth immediately following a purge. As Christy mentioned, *"I think harm reduction is great. I love—because I purge. And, if it weren't for harm reduction, I think I would've fucked my teeth up so much more than I have."* Harm reduction provides resources for individuals who have an eating disorder, but cannot or will not recover, to stay safe and informed. Despite benefits, harm reduction resources are treated differently across eating disorder spaces online. While someone communities freely permit them, others such as a new incarnation of r/ProED, have active moderation teams dedicated to removing posts related to tips or advice and carefully overseeing content related to harm reduction. These community-led practices differ from their historical precedent, in which harm reduction was not liable for removal or modification at the discretion of a moderation team. This example illustrates an easy to miss point—that harm reduction, previously unregulated, is now subject to new practices of community-led moderation, which stand to potentially restrict this beneficial form of content, due to interactions with past and current platform practices of moderation. Here, then, we see how platform moderation contributes additional labor to ingroup moderators [29, 118] and interacts with community-led moderation to shape certain versions of eating disorders online.

5 DISCUSSION

In the following sections, we turn to conformity as a way of understanding content moderation and the way it contributes to broad social and structural effects on marginalized groups. We discuss how moderation works to establish and enforce conformity, particularly among people with eating disorders, and its consequences for members of marginalized groups. In contrast to conformity, we then discuss how platforms can design for a multiplicity of illness experiences online.

5.1 Content Moderation as Enforcing an Order of the Normal

People with eating disorders have historically been subjected to processes of conformity that aim to dictate overarching norms and values, particularly with respect to the enforcement of certain ideas of body image and mental illness. These processes are not merely projected onto particular groups

or communities (e.g., through guidelines, codes of conduct, or diagnostic manuals), but in practice come to be enacted through unfolding relations between varied actors and the sociotechnical structures in which they operate. Take, for example, the seemingly objective indicator of energy in food, calories. For people with anorexia, calories become entangled in a web of sociotechnical structures and normative value systems. The health profession recommends a staged process of weight-gain, with monitoring based on calorie counts [42]. Thus, a version of normal, in this case based on body weight and energy consumption, is enforced through a program of monitoring and control. Moreover, regulations, standards, and technological innovations help in this tracking of calorific intake and exertion [88]. The processes of conformity are then enacted through varied technoscientific structures, affording and enforcing a distinctive regime of norms.

Following from the materials presented above, we argue that social media platforms play a part in the wider sociotechnical processes of conformity. Specifically, across social media platforms, conformity to versions of mental illness and body image is established and enforced through platform features and the capacities for interaction that are afforded through them (e.g., commenting; labeling value through 'likes'; content promotion and demotion via features, such as 'upvotes' and 'downvotes' and algorithms that prioritize content). Our claim is that through practices of content moderation these platforms are, in effect, enforcing an *order of normal*. With respect to our findings, it is altogether too easy to attend to specific instances within our data, such as the particular wordings of a comment, the reporting of a specific post, or the deletion of an individual's account. However, by shifting our attention to what is happening across these cases and across an ecosystem of online spaces, we can see the structural forces at play.

Consider the reporting features on social media platforms. By approaching their design and use in terms of the wider structural practices of content moderation, we get a clearer picture of how conformity and an order of normal are enforced. As we have seen, the threat of being reported regulates online behavior; in particular, what people are willing to say about eating disorders and their own actions and beliefs. Critically, the power of moderation in this context is not in the reporting *per se*, but in its perpetual threat. Reminiscent of Foucault's well-rehearsed reading of the panopticon [38], the punishment is no longer the prime means of control. Rather, it is surveillance. We see this being particularly effective on a platform like Instagram, which provides reporting features to anyone on the internet, whether or not they hold an account. In making a report, anonymity is preserved, allowing for control without accountability to consequences. This particular type of reporting opens up groups, especially those at the margins, to the judgement and force of a pervasive and invisible surveillance. The Panoptic qualities of a platform (see Wood [119]), or more broadly its structural configuration designed to support content moderation, controls users and, in the case of eating disorders, regulates content so that it adheres to a norm. What the structural mechanisms of content moderation serve to do, then, is actively delineate the boundaries of what is acceptable within particular online spaces. They come to constitute a 'structural machinery' that sanctions some bodies and forms of mental illness, while simultaneously casting others as other-than-normal or deviant. Classifications of wrongness and deviance are amplified by practices of moderation that target certain versions of eating disorders, at times removing them from social media platforms and, therefore, the ability to participate in constituting versions of mental illness and body image online.

Key to this structural framing is that individual instances of moderation, moderation that occurs within and among community groups, and distinctive interactive features of the platform must be understood together. In our findings, we captured online moderation practices at different 'levels', including those of the ingroup community and those operating at the level of the platform. We want to make clear that considering these together as a whole, rather than distinct from one another, allows a stronger foundation for understanding what content moderation across social

media platforms is doing. Reddit's form of community quarantine offers a helpful example. Once a community has been quarantined by the platform, constraints are enforced using specific technical configurations. Here, a limbo status is conferred upon the community so that content from the quarantined subreddit does not appear in platform search results or on non-subscription feeds, such as r/Popular. This greatly reduces the connectivity of the quarantined community from the rest of Reddit, which, in turn, can impact its participation and growth. Though content is not always removed from the community or the platform, the subreddit's status is demoted, and accordingly, so are the norms and values it espouses. This speaks to ways in which moderation and its consequences on eating disorder groups are deeply entangled with the structures of a platform. The norms surrounding eating disorders, mental illness, and body image are enmeshed with how content moderation is designed and programmed into a platform.

Consequences of conformity clearly resonate with feminist and biopolitical accounts of bodies [10, 13, 41] and, in particular, the ways social media platforms exert structural forms of control on the ordering of bodies [26]. Current platform practices, as well as research and design suggestions that call for additional, albeit different, forms of moderation and intervention, may result in new aggregates of human and technical control that work to establish and enforce existing normativities [18, 25, 59, 61, 89]. This applies, even, to recent research on content moderation that suggests a shift toward an educational paradigm rather than a punitive one [89]. Underlying this and other recommendations is the assumption that social media users should internalize, rather than question or counter, mechanisms of control. We argue platforms and researchers should aim to support individuals with eating disorders, rather than impose an order of normal that further marginalizes and subjugates experiences that need to be valued and expressed.

5.2 Consequences of Conformity for Marginalized Groups

Though social media platforms are private companies, they have arguably become public spaces [65, 101], particularly when examined collectively as an ecosystem. Specific moderation practices on any one platform might not pose a problem in isolation. However, the ways in which platforms operate in similar ways with respect to content related to eating disorders contributes to systemic discriminatory practices and displacement of individuals on the margins. By addressing the consequences of conformity, as we do here, we contribute to a developing body of work that examines how the sociotechnical machinery of platforms and algorithms (e.g., on social media and elsewhere) exclude non-normative identities and forms of expression and interaction [5, 6, 51, 54].

Posts about mental illness on social media are a type of illness narrative [33]. These narratives provide opportunities for individuals to work through, reflect on, and communicate the subjective experience of being ill [39, 67]. Though illness narratives frequently arise in the context of other experiences, including cancer and chronic illness [108, 120], tellings of mental illness, particularly those sharing experiences that may be "ugly", differ in that their likelihood of moderation and platform removal is high. This is troubling. All narratives of mental illness shared by those living the experience are valid and deserve to be voiced. However, the impact of moderation is such that individuals may respond by concealing the full extent of their eating disorder on social media. This poses a problem, particularly for the many individuals for whom social media platforms may be the only spaces in which they feel comfortable disclosing and discussing their experiences. Without those spaces, and without others elsewhere in their lives, these individuals are at risk for psychological consequences related to hiding stigmatized experiences [94].

Processes of conformity also operate to displace individuals and communities. Displacement may be likened to a form of digital gentrification [79], in which marginalized groups are forcibly removed from platforms to benefit a majority. Given how platforms generate revenue, displacement

of individuals with eating disorders might be conducted to present a vision of an advertisement-friendly social media. However, displacement in this context goes beyond considerations of revenue production. It creates inequalities with respect to content production and which voices are permitted online. Displacement often occurs alongside moderation and simultaneous provisions of helpline resources. Though these resources may provide valuable and informative assets for individuals with eating disorders, as well as their family members and friends, they do not—and cannot—replace support networks and opportunities for disclosure. Further, when support resources are limited, such as to the NEDA helpline, they present a bounded interpretation of life with an eating disorder. These constraints ignore personal histories and experiences with eating disorders, including recovery, relapse, and management, demonstrating how a blanket solution (e.g., providing the same resources for everyone) may ultimately fail many.

In considering regulation and subsequent responses, it is vital for platforms to acknowledge the labor they create for individuals with eating disorders, a group of people who face marginalization and stigma in online and offline spaces [82, 107]. As other reports have shown [93], social media platforms can negatively impact marginalized groups (e.g., Rohingya people, Native Americans, Black Americans) through practices of content moderation. Though we have also found consequences, we additionally show how individuals can use certain platform features to subvert control and moderation. These forms of resistance, which others have aligned with practices of civil disobedience [89], share their spirit and their histories with other forms of social activism, including those related to Mad Pride and MeToo [75, 86]. However, resistance is itself a burden. It requires individuals with marginal status to go beyond typical platform interactions just to enjoy the same access.

5.3 Supporting a Multiplicity of Eating Disorders

Important for rethinking design are cases where we see moderation and conformity operating in constructive ways. Several of our participants discussed practices of moderation that resulted in a conformity of mental illness and body image tailored to respect the standards of their ingroup community (e.g., forum organization, content warnings). Though these processes of moderation and conformity operate in a similar way to other online platforms and social media sites (e.g., through sociotechnical relations), they're practiced in ways that support diverse experiences, rather than restrictive ones. This highlights promising possibilities for supporting the multiplicities inherent in groups or communities of individuals with eating disorders [87].

5.3.1 Shifting Away from Deviance. Much of the experience of eating disorders is not to be liked. It can be ugly and painful—recovery included. However, rather than casting content as non-normative or deviant (i.e., such as through its removal or by other technical configurations that set it apart), social media platforms, should reconsider the ways in which illness, of any type, is addressed and moderated online. As we describe above, illness narratives can be productive ways to document and share the experience of being ill. Yet, these personal narratives are at risk when they do not adhere to the order of normal enforced on social media platforms and online communities. Rather than constraining experience to normative or not, such as through practices of content moderation disproportionately impacting non-normative types of content, platforms should reconsider ways to rework interactions surrounding narratives of illness. To this effect, rather than reporting features or educating those who have been subject to moderation [59, 89], platforms could provide educational resources to others (i.e., people reporting content) about the experience of living with mental illness and importance of disclosing illness narratives.

5.3.2 Coexisting through Strategic Content Practices. Designing for multiplicity also means mindfully attending to the ways that different types of people coexist with one another. Certain ways of

posting about eating disorders, including experiences with illness and recovery, can be triggering or upsetting for others. Content warnings, as our informants described, are one way to coexist and safely access online spaces. This is, of course, with the caveat that content warnings are not universal. Platforms, therefore, have an opportunity to support design in this context that permits expression and disclosure, as well as safety and access. For example, Instagram has a content warning feature which blurs an image until a viewer decides to click on it. However, this content warning is platform-applied. Our data suggests that individuals with eating disorders seem likely to appropriate such features as a way of online self-preservation and community sustainment. As part of reworking the notion of content warnings, we should consider how platform features could help people be more aware of their potential audiences and create mindfulness around how varied others may interpret certain content. For example, for some, simply having numbers (e.g., calorie counts) in a post can be triggering. Through new and improved mechanisms for self-moderation, platforms could make available technical affordances that maintain freedom of expression *and* help individuals navigate content.

5.3.3 Reconfiguring Power Dynamics. Another way that platforms and online communities can move toward a more equitable and just experience online is to shift power dynamics embedded within content moderation features and practices. In particular, platform features could give more agency to the individual who has been reported rather than the individual who has done the reporting (i.e., the reporter)—or at least aim to strike a balance. Not being able to face another user or computational actor who played a role in content moderation, or even know how or what happened when content was reported or erased, contributes to the marginalization that individuals with eating disorders face on a day-to-day basis. Further, the emotional burden and labor of restoring activity online, including finding a space to exist, is shouldered by those who may need support the most. Productive design changes may include increasing transparency with respect to moderation and its motivation; temporarily archiving an individual's account or content during a process of deliberation (i.e., rather than erasure); and turning moderated content over to individuals to restore their control over their personal data and preserve an important part of their illness narrative, which may be useful now or in the future for self-reflection and growth.

6 CONCLUSION

Practices of content moderation are integral to what social media platforms do. However, they are far from perfect and increasingly difficult to get right. Despite good intentions, practices of moderation have consequences for individuals with eating disorders and other members of marginalized groups, including loss of personal content and community support, and labor associated with practices of resistance. In this paper, we examine the experience of content moderation and how, in particular, practices of content moderation (e.g., content removal, quarantine, support resources) work to establish and enforce a conformity to mental illness and body image on social media. We argue that processes of conformity as reproduced through sociotechnical structures afforded by platforms work to exclude people with eating disorders and other non-normative identities and behaviors. Rather than design for restrictive content moderation practices, we suggest that platforms consider supporting a diversity of eating disorder and illness experiences by designing for multiplicity.

REFERENCES

[1] Nazanin Andalibi. 2019. What Happens After Disclosing Stigmatized Experiences on Identified Social Media: Individual, Dyadic, and Social/Network Outcomes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 137.

[2] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative*

Work and Social Computing (CSCW '17). ACM, New York, NY, USA, 1485–1500. https://doi.org/10.1145/2998181.2998243

[3] Monica Anderson, Skye Toor, Lee Rainie, and Aaron Smith. 2018. Activism in the Social Media Age. *Pew Research Center* (11 July 2018). https://www.pewinternet.org/2018/07/11/activism-in-the-social-media-age/

[4] Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. 2015. Social sharing of emotions on Facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, 154–164.

[5] Rena Bivens. 2017. The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society* 19, 6 (2017), 880–898.

[6] Rena Bivens and Oliver L Haimson. 2016. Baking gender into social media design: How platforms shape categories for users and advertisers. *Social Media+ Society* 2, 4 (2016), 2056305116672486.

[7] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 24:1–24:19. https://doi.org/10.1145/3134659

[8] Dina L. G. Borzekowski, Summer Schenk, Jenny L. Wilson, and Rebecka Peebles. 2010. e-Ana and e-Mia: A Content Analysis of Pro-Eating Disorder Web Sites. *American Journal of Public Health* 100, 8 (Aug. 2010), 1526–1534. https://doi.org/10.2105/AJPH.2009.172700

[9] Leanne Bowler, Eleanor Mattern, Wei Jeng, Jung Sun Oh, and Daqing He. 2013. I know what you are going through: answers to informational questions about eating disorders in Yahoo! Answers: a qualitative study. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*. American Society for Information Science, 6.

[10] Rosi Braidotti. 2000. Teratologies. In *Deleuze and Feminist Theory*, Ian Buchanan and Claire Colebrook (Eds.). Edinburgh University Press, Edinburgh, 156–172. https://edinburghuniversitypress.com/book-deleuze-and-feminist-theory.html

[11] Amy Bruckman. 2006. Teaching students to study online communities ethically. *Journal of Information Ethics* 15, 2 (2006), 82.

[12] Eleanor R. Burgess, Kathryn E. Ringland, Jennifer Nicholas, Ashley A. Knapp, Jordan Eschler, David C. Mohr, and Madhu C. Reddy. 2019. "I think people are powerful": The sociality of individuals managing depression. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov 2019), 41:1–41:29. https://doi.org/10.1145/3359143

[13] Judith Butler. 1993. *Bodies That Matter: On the Discursive Limits of 'Sex'*. Routledge, New York.

[14] Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina J Costello, Nina Kaiser, Elizabeth S Cahn, Ellen E Fitzsimmons-Craft, and Denise E Wilfley. 2019. "I just want to be skinny.": A content analysis of tweets expressing eating disorder symptoms. *PloS one* 14, 1 (2019), e0207506.

[15] Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, and David A. Shamma. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3213–3226. https://doi.org/10.1145/3025453.3025985

[16] Stevie Chancellor, Zhiyuan (Jerry) Lin, and Munmun De Choudhury. 2016. "This Post Will Just Get Taken Down": Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1157–1162. https://doi.org/10.1145/2858036.2858248

[17] Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016. Recovery Amid Pro-Anorexia: Analysis of Recovery in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2111–2123. https://doi.org/10.1145/2858036.2858246

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

[19] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 31.

[20] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 32:1–32:25. https://doi.org/10.1145/3274301

[21] Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. *arXiv preprint arXiv:1909.01362* (2019).

[22] Kathy Charmaz. 2014. *Constructing Grounded Theory*. SAGE.

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

[24] Sky Croeser. 2016. Thinking Beyond 'Free Speech' in Responding to Online Harassment. *Ada: a journal of new media, gender, and culture* 10 (2016), online–online.

[25] Munmun De Choudhury. 2015. Anorexia on Tumblr: A Characterization Study. In *Proceedings of the 5th International Conference on Digital Health 2015 (DH '15)*. ACM, New York, NY, USA, 43–50. https://doi.org/10.1145/2750511.2750515

[26] Tisha Dejmanee. 2013. Bodies of technology : performative flesh, pleasure and subversion in cyberspace. *Gender Questions* 1, 1 (Jan. 2013), 3–17. https://journals.co.za/content/genderq/1/1/EJC167619

[27] Michael A. DeVito, Ashley Marie Walker, and Jeremy Birnholtz. 2018. 'Too Gay for Facebook': Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 44:1–44:23. https://doi.org/10.1145/3274313

[28] Jill P Dimond, Michaelanne Dye, Daphne LaRose, and Amy S Bruckman. 2013. Hollaback!: the role of storytelling online in a social movement organization. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 477–490.

[29] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 142.

[30] Maeve Duggan. 2017. Online harassment 2017. (2017).

[31] Brianna Dym, Jed R Brubaker, Casey Fiesler, and Bryan Semaan. 2019. "Coming Out Okay": Community Narratives for LGBTQ Identity Recovery Work. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 29. https://doi.org/10.1145/3359256

[32] Jessica L. Feuston and Anne Marie Piper. 2018. Beyond the Coded Gaze: Analyzing Expression of Mental Health and Illness on Instagram. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 51:1–51:21. https://doi.org/10.1145/3274320

[33] Jessica L. Feuston and Anne Marie Piper. 2019. Everyday Experiences: Small Stories and Mental Illness on Instagram. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 265:1–265:14. https://doi.org/10.1145/3290605.3300495 event-place: Glasgow, Scotland Uk.

[34] Casey Fiesler, Michaelanne Dye, Jessica L Feuston, Chaya Hiruncharoenvate, Clayton J Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S Bruckman, Munmun De Choudhury, et al. 2017. What (or who) is public?: Privacy settings and social media content sharing. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 567–580.

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

[36] Rachel A Fleming-May and Laura E Miller. 2010. I'm scared to look but I'm dying to know: information seeking and sharing on Pro-Ana weblogs. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem-Volume 47*. American Society for Information Science, 61.

[37] Michel Foucault. 2003. *Madness and civilization*. Routledge.

[38] Michel Foucault. 2012. *Discipline and punish: The birth of the prison*. Vintage.

[39] Arthur W. Frank. [n. d.]. *The Wounded Storyteller*. https://www.press.uchicago.edu/ucp/books/book/chicago/W/bo14674212.html

[40] Mary Anne Franks. [n. d.]. The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment? *Knight First Amendment Institute at Columbia University* ([n. d.]). https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment

[41] Christian Fuchs. 2013. *Internet and Surveillance: The Challenges of Web 2.0 and Social Media*. Routledge.

[42] Andrea K Garber, Susan M Sawyer, Neville H Golden, Angela S Guarda, Debra K Katzman, Michael R Kohn, Daniel Le Grange, Sloane Madden, Melissa Whitelaw, and Graham W Redgrave. 2016. A systematic review of approaches to refeeding in patients with anorexia nervosa. *International Journal of Eating Disorders* 49, 3 (2016), 293–310.

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

[44] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

[45] Debbie Ging and Sarah Garvey. 2018. 'Written in these scars are the stories I can't explain': A content analysis of pro-ana and thinspiration image sharing on Instagram. *New Media & Society* 20, 3 (2018), 1181–1200.

[46] Eric Goldman. 2010. Unregulating Online Harassment. *Denver University Law Review Online* 57 (2010), 59.

[47] Mary L Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books, New York.

Submitted 16 Oct 2019. Published May 2020

[48] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* (2015), 42–109. https://heinonline.org/HOL/P?h=hein.journals/yjolt17&i=42

[49] Oliver L Haimson, Jed R Brubaker, Lynn Dombrowski, and Gillian R Hayes. 2015. Disclosure, stress, and support during gender transition on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1176–1190.

[50] Oliver L Haimson, Jed R Brubaker, Lynn Dombrowski, and Gillian R Hayes. 2016. Digital footprints and changing networks during online identity transitions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2895–2907.

[51] Oliver L Haimson and Anna Lauren Hoffmann. 2016. Constructing and enforcing" authentic" identity online: Facebook, real names, and non-normative identities. *First Monday* 21, 6 (2016).

[52] Oliver L Haimson, Kathryn E Ringland, and Gillian R Hayes. 2015. Marginalized populations and research ethics online. In *CSCW Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World (CSCW Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World)*, Vol. 5.

[53] Max Halupka. 2014. Clicktivism: A systematic heuristic. *Policy & Internet* 6, 2 (2014), 115–132.

[54] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 8.

[55] Libby Hemphill and Andrew J Roback. 2014. Tweet acts: how constituents lobby congress via Twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 1200–1210.

[56] Dorothy Howard and Lilly Irani. 2019. Ways of Knowing When Research Subjects Care. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 97.

[57] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 611–620.

[58] Grace YoungJoo Jeon, Nicole B Ellison, Bernie Hogan, and Christine Greenhow. 2016. First-generation students and college: The role of Facebook networks as information sources. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 887–899.

[59] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *ACM Trans. Comput.-Hum. Interact.* 3, CSCW (Nov. 2019), 33. https://doi.org/10.1145/3359294

[60] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5 (July 2019), 31:1–31:35. https://doi.org/10.1145/3338243

[61] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *ACM Trans. Comput.-Hum. Interact.* 3, CSCW (Nov. 2019), 27. https://doi.org/10.1145/3359252

[62] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 2, CSCW (March 2018), 12:1–12:33. https://doi.org/10.1145/3185593

[63] Jialun "Aaron" Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 23.

[64] Adrienne S Juarascio, Amber Shoaib, and C Alix Timko. 2010. Pro-eating disorder communities on social networking sites: a content analysis. *Eating disorders* 18, 5 (2010), 393–407.

[65] Jeffrey S Juris. 2012. Reflections on# Occupy Everywhere: Social media, public space, and emerging logics of aggregation. *American Ethnologist* 39, 2 (2012), 259–279.

[66] Charles Kiene, Jialun "Aaron" Jiang, and Benjamin Mako Hill. 2019. Technological Frames and User Innovation: Exploring Technological Change in Community Moderation Teams. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 23.

[67] Arthur Kleinman. 1989. *The Illness Narratives: Suffering, Healing, And The Human Condition* (reprint edition ed.). Basic Books, New York.

[68] Priya Kumar and Sarita Schoenebeck. 2015. The modern day baby book: Enacting good mothering and stewarding privacy on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1302–1312.

[69] Lakesha Lafayett. 2017. The Problematic Whitewashing of Eating Disorder Recovery. https://www.nationaleatingdisorders.org/blog/problematic-whitewashing-eating-disorder-recovery

[70] Andrea LaMarre and Carla Rice. 2017. Hashtag Recovery: #Eating Disorder Recovery on Instagram. *Social Sciences* 6, 3 (June 2017), 68. https://doi.org/10.3390/socsci6030068

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

[72] Cliff A.C. Lampe, Erik Johnston, and Paul Resnick. 2007. Follow the Reader: Filtering Comments on Slashdot. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1253–1262. https://doi.org/10.1145/1240624.1240815 event-place: San Jose, California, USA.

[73] landoflobsters. 2019. Changes to Our Policy Against Bullying and Harassment. *Reddit* (30 September 2019). https://www.reddit.com/r/announcements/comments/dbf9nj/changes_to_our_policy_against_bullying_and/

[74] Yu-Hao Lee and Gary Hsieh. 2013. Does slacktivism hurt activism?: the effects of moral balancing and consistency in online activism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 811–820.

[75] Brenda A LeFrançois, Robert Menzies, and Geoffrey Reaume. 2013. *Mad matters: A critical reader in Canadian mad studies*. Canadian Scholars' Press.

[76] A Lenhart, M Ybarra, K Zickuhr, and M Price-Feeney. 2016. Online Harassment, Digital Abuse, and Cyberstalking in America. Data & Society Research Institute. *Center for Innovative Public Health Research. Retrieved from: https://datasociety. net/pubs/oh/Online_Harassment_2016. pdf* (2016).

[77] Philip LeVine and Ron Scollon. 2004. Multimodal Discourse Analysis as the Confluence of Discourse and Technology. In *Discourse and technology: Multimodal discourse analysis*, Philip LeVine and Ron Scollon (Eds.). Georgetown University Press, 1–6.

[78] Stephen P Lewis and Alexis E Arbuthnott. 2012. Searching for thinspiration: the nature of internet searches for pro-eating disorder websites. *Cyberpsychology, Behavior, and Social Networking* 15, 4 (2012), 200–204.

[79] Jessa Lingel. 2019. The gentrification of the internet. http://culturedigitally.org/2019/03/the-gentrification-of-the-internet/

[80] Deanna Linville, Tiffany Brown, Katrina Sturm, and Tori McDougal. 2012. Eating disorders and social support: perspectives of recovered individuals. *Eating Disorders* 20, 3 (2012), 216–231.

[81] Fannie Liu, Denae Ford, Chris Parnin, and Laura Dabbish. 2017. Selfies as social movements: Influences on participation and perceived impact on stereotypes. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 72.

[82] James D. Livingston and Jennifer E. Boyd. 2010. Correlates and consequences of internalized stigma for people living with mental illness: A systematic review and meta-analysis. *Social Science & Medicine* 71, 12 (Dec. 2010), 2150–2161. https://doi.org/10.1016/j.socscimed.2010.09.030

[83] Ben Marder, Adam Joinson, Avi Shankar, and David Houghton. 2016. The extended 'chilling' effect of Facebook: The cold reality of ubiquitous social networking. *Computers in Human Behavior* 60 (2016), 582–592.

[84] Aiden McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Cranefield. 2016. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. *ICIS 2016 Proceedings* (Dec. 2016). https://aisel.aisnet.org/icis2016/SocialMedia/Presentations/23

[85] Bharat Mehra, Cecelia Merkel, and Ann Peterson Bishop. 2004. The internet for empowerment of minority and marginalized users. *New media & society* 6, 6 (2004), 781–802.

[86] Kaitlynn Mendes, Jessica Ringrose, and Jessalynn Keller. 2018. #MeToo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies* 25, 2 (May 2018), 236–246. https://doi.org/10.1177/1350506818765318

[87] Annemarie Mol. 2002. *The body multiple: Ontology in medical practice*. Duke University Press.

[88] Annemarie Mol. 2013. Mind your plate! The ontonorms of Dutch dieting. *Social studies of science* 43, 3 (2013), 379–396.

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

[90] Jung Sun Oh, Daqing He, Wei Jeng, Eleanor Mattern, and Leanne Bowler. 2013. Linguistic characteristics of eating disorder questions on Yahoo! Answers-content, style, and emotion. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*. American Society for Information Science, 87.

[91] Kathleen O'Leary, Arpita Bhattacharya, Sean A Munson, Jacob O Wobbrock, and Wanda Pratt. 2017. Design opportunities for mental health peer support technologies. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1470–1484.

[92] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, 994–1009.

[93] Onlinecensorship.org. 2019. Offline-Online. https://onlinecensorship.org/content/infographics Retrieved September 20, 2019.

[94] John E Pachankis. 2007. The psychological implications of concealing a stigma: A cognitive-affective-behavioral model. *Psychological bulletin* 133, 2 (2007), 328.

Submitted 16 Oct 2019. Published May 2020

[95] Jessica A. Pater, Oliver L. Haimson, Nazanin Andalibi, and Elizabeth D. Mynatt. 2016. "Hunger Hurts but Starving Works": Characterizing the Presentation of Eating Disorders Online. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1185–1200. https://doi.org/10.1145/2818048.2820030

[96] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work (GROUP '16)*. ACM, New York, NY, USA, 369–374. https://doi.org/10.1145/2957276.2957297 event-place: Sanibel Island, Florida, USA.

[97] Jessica A. Pater, Lauren E. Reining, Andrew D. Miller, Tammy Toscos, and Elizabeth D. Mynatt. 2019. "Notjustgirls": Exploring Male-related Eating Disordered Content Across Social Media Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 651:1–651:13. https://doi.org/10.1145/3290605.3300881 event-place: Glasgow, Scotland Uk.

[98] Jon Penney. 2017. Internet surveillance, regulation, and chilling effects online: A comparative case study. *Internet Policy Review* (2017).

[99] Shruti Phadke, Jonathan Lloyd, James Hawdon, Mattia Samory, and Tanushree Mitra. 2018. Framing Hate with Hate Frames: Designing the Codebook. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 201–204.

[100] Radha Iyengar Plumb. 2019. An Independent Report on How We Measure Content Moderation. *Facebook Newsroom* (23 May 2019). https://newsroom.fb.com/news/2019/05/dtag-report/

[101] Thomas Poell and José Van Dijck. 2016. Constructing public space: Global perspectives on social media and popular contestationâĂŤIntroduction. *International Journal of Communication* 10 (2016), 9.

[102] Rebecca Puhl and Young Suh. 2015. Stigma and eating and weight disorders. *Current psychiatry reports* 17, 3 (2015), 10.

[103] Bryce J Renninger. 2015. "Where I can be myself... where I can speak my mind": Networked counterpublics in a polymedia environment. *New Media & Society* 17, 9 (2015), 1513–1529.

[104] Sarah Roberts. 2016. Digital Refuse: Canadian Garbage, Commercial Content Moderation and the Global Circulation of Social Media's Waste. *Wi: Journal of Mobile Media* (Jan. 2016). https://ir.lib.uwo.ca/commpub/14

[105] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

[106] James P Roehrig and Carmen P McLean. 2010. A comparison of stigma toward eating disorders versus depression. *International Journal of Eating Disorders* 43, 7 (2010), 671–674.

[107] Nicolas Rüsch, Matthias C. Angermeyer, and Patrick W. Corrigan. 2005. Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *European Psychiatry* 20, 8 (Dec. 2005), 529–539. https://doi.org/10.1016/j.eurpsy.2005.04.004

[108] Shruti Sannon, Elizabeth L. Murnane, Natalya N. Bazarova, and Geri Gay. 2019. "I Was Really, Really Nervous Posting It": Communicating About Invisible Chronic Illnesses Across Social Media Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 353:1–353:13. https://doi.org/10.1145/3290605.3300583 event-place: Glasgow, Scotland Uk.

[109] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 155:1–155:27. https://doi.org/10.1145/3274424

[110] Andrew Scull. 2015. *Madness in Civilization: A Cultural History of Insanity, from the Bible to Freud, from the Madhouse to Modern Medicine*. Princeton University Press. 432 pages.

[111] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277 event-place: Portland, Oregon, USA.

[112] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (July 2019), 1417–1443. https://doi.org/10.1177/1461444818821316

[113] Leslie Regan Shade. 2003. Weborexics: The ethical issues surrounding pro-ana websites. (2003).

[114] Walter Vandereycken and Ina Van Humbeeck. 2008. Denial and concealment of eating disorders: a retrospective survey. *European Eating Disorders Review* 16, 2 (2008), 109–114. https://doi.org/10.1002/erv.857

[115] Aditya Vashistha, Abhinav Garg, Richard Anderson, and Agha Ali Raza. 2019. Threats, Abuses, Flirting, and Blackmail: Gender Inequity in Social Media Voice Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 72.

[116] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1231–1245.

[117] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 91–100.

[118] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 160.

[119] David Wood. 2003. Foucault and panopticism revisited. *Surveillance & Society* 1, 3 (2003), 234–239.

[120] Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The Channel Matters: Self-disclosure, Reciprocity and Social Support in Online Cancer Support Groups. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 31:1–31:15. https://doi.org/10.1145/3290605.3300261 event-place: Glasgow, Scotland Uk.

Submitted 16 Oct 2019. Published May 2020