Beyond Initial Removal: Lasting Impacts of Discriminatory Content Moderation to Marginalized Creators on Instagram

YIM REGISTER, University of Washington, USA
IZZI GRASSO, University of Washington, USA
LAUREN N. WEINGARTEN, University of Washington, USA
LILITH FURY, University of Washington, USA
CONSTANZA ELIANA CHINEA, University of Washington, USA
TUCK J. MALLOY, University of Washington, USA
EMMA S. SPIRO, University of Washington, USA

Recent work has demonstrated how content moderation practices on social media may unfairly affect marginalized individuals, for example by censoring women's bodies and misidentifying reclaimed terms as hate speech. This study documents and explores the direct experiences of marginalized creators who have been impacted by discriminatory content moderation on Instagram. Collaborating with our participants for over a year, we contribute five co-constructed narratives of discriminatory content moderation from advocates in trauma-informed care, LGBTQ+ sex education, anti-racism education, and beauty and body politics. In sharing these detailed personal accounts, not only do we shed light on their experiences with being blocked, banned, or deleted unfairly, but we delve deeper into the lasting impacts of these experiences to their livelihoods and mental health. Reflecting on their stories, we observe that content moderation on social media is deeply entangled with the situated experiences of offline discrimination. As such, we document how each participant experiences moderation through the lens of their often intersectional identities. Using participatory research methods, we collectively strategize ways to learn from these individual accounts and resist discriminatory content moderation, as well as imagine possibilities for repair and accountability.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; HCI design and evaluation methods; • **Social and professional topics** → **Censorship**.

Additional Key Words and Phrases: content moderation; instagram; social media; LGBTQ; race; gender; marginalization; hate speech; digital activism; shadowban; algorithm bias

ACM Reference Format:
Yim Register, Izzi Grasso, Lauren N. Weingarten, Lilith Fury, Constanza Eliana Chinea, Tuck J. Malloy, and Emma S. Spiro. 2024. Beyond Initial Removal: Lasting Impacts of Discriminatory Content Moderation to Marginalized Creators on Instagram. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 23 (April 2024), 28 pages. https://doi.org/10.1145/3637300

1 INTRODUCTION

Imagine you have spent years developing an Instagram page in support of survivors of sexual violence. You post recovery advice, mental health resources, and other timely content. The page is a community hub for peer support. You've built up over 20K followers, dedicating your time and energy to the page

Authors' addresses: Yim Register, University of Washington, Seattle, WA, USA; Izzi Grasso, University of Washington, Seattle, WA, USA; Lauren N. Weingarten, University of Washington, Seattle, WA, USA; Lilith Fury, University of Washington, Seattle, WA, USA; Constanza Eliana Chinea, University of Washington, Seattle, WA, USA; Tuck J. Malloy, University of Washington, Seattle, WA, USA; Emma S. Spiro, University of Washington, Seattle, WA, USA.

ACM 2573-0142/2024/4-ART23
https://doi.org/10.1145/3637300

advocating against sexual violence. One day you log on to Instagram and there is nothing there. Your account is completely gone. No explanation. No information. You are completely in the dark.

This type of experience is common for advocates on Instagram (IG). They frequently face unclear and unreliable content moderation, often without access to a valid appeals process [70]. While we know that discriminatory content moderation is occurring, and prior work has documented specific cases [45], we know less about how the *experience* of being moderated impacts user's daily lives, access to community, and sense of self. Through participatory methods and documenting of experiential stories co-constructed alongside marginalized creators, we offer rich and deeply personal insight into the downstream effects of content moderation practices, with a focus on the lasting impacts to user wellbeing. In five case studies, we demonstrate the breadth of "content moderation gray areas" [45] as well as the situated experiences of marginalized creators; we demonstrate how moderation interacts with their identities, privileges, access, and mental health. Our findings illustrate the lasting impacts of discriminatory content moderation decisions, as told by creators themselves. Each story points to avenues for much-needed future research studying the unintended consequences of recommendations and polices of moderation online.

Scholars agree that Instagram has transformed into much more than a social networking site to connect with friends. In particular, many accounts are social advocacy and/or educational pages led by marginalized creators. For the purposes of this work, we view marginalization as a form of exclusion, oppression, and systemic lack of access due to perceived and constructed social difference in categories such as race, class, gender, and/or ability [19, 20, 42]. Systems of power and domination oppress across social difference in ways that are interlocking and multidimensional; and those who are categorized as different from the norm across any of these systems are marginalized, or systematically excluded – via policy, opportunity, or social enforcement. This work in particular explores experiences of advocates in trauma-informed care, LGBTQ+ sex education, anti-racism education, and beauty and body politics. Recent scholarship has demonstrated that marginalized users are often disproportionately affected by content moderation [4, 28, 37, 39, 41, 45, 100]; either directly via what is not allowed in Community Guidelines or by how those guidelines are deferentially enforced. For example, "female nipples" are not allowed to be visible according to IG's Community Guidelines. In other cases, LGBTQ+ social media users are mistakenly labeled as using hate speech when speaking about their own identities, e.g. using the term "bitch" as a term of endearment, referring to one's own trans identities, or using other reclaimed terms [27, 45]. Sometimes marginalized users re-post comments and messages they have received that contain hate speech, harassment, and bullying – these "callout" posts sometimes get taken down and yet the original comments do not, a stark failure of the system [70, 100]. Moderation can interact with one's marginalized identities; for example, Haimson et al. [45] conducted surveys to characterize and quantify disproportionate removals of content across Facebook, Twitter, and Instagram. They found that political conservatives, transgender people, and Black people experienced content and account removals more often than others. For Black and/or Trans social media users, the content removed typically had to do with them expressing their marginalized identities.

Human-computer interaction (HCI), computational social science, design, and misinformation research (just to name a few) all contend with issues of content moderation on social media. In earnest attempts to avoid harm, research in these spaces often recommends uniform, punitive actions, such as restricting all nudity or quickly removing accounts found responsible for repeatedly spreading harmful content (as deemed by the platform) [52]. What we know little about, and is often ignored in these recommendations, are the downstream harms of these policies and recommendations. CSCW has successfully engaged in recommendations on providing more transparency [50], envisioning moderation beyond punishment and as opportunity for restorative justice [57], and the importance of discourse and education around community guidelines [49]. We contribute

further insight into the effects of content moderation policies to vulnerable groups – furthering the scholarship in this area that seeks to engage with the difficulty of designing effective content moderation practices.

We define discriminatory content moderation as the disparate moderation of marginalized users by social media platforms. This could include, but is not limited to, increasing reach and engagement to users that hold normative identities in a way that is inaccessible to marginalized users, deleting or reducing reach and engagement of the content of marginalized users that are not violating Community Guidelines, and maintaining platform policies or guidelines that disadvantage marginalized users. This study contributes five situated narratives of content moderation, along with an important call to action for how we, as a research community, might structure priorities in order to mitigate harm to marginalized communities. Through participatory research and co-construction of narratives with our participants, this work aims to make space for these creators to accurately represent their *own* stories. Over the course of a year, we conducted interviews and follow-up narrative editing sessions with five creators, each with membership in and knowledge of specific marginalized communities, who have experienced discriminatory content moderation on IG. Each creator was recruited based on specific expertise which we elaborate on in Section 3.2. Their stories illustrate the enmeshment of online content moderation and offline discrimination and harm, with content moderation often reproducing larger systems of oppression in unseen and automated ways. In both our methods and discussion, we draw from Chen et al. [16]'s framework for understanding technology experiences through a trauma-informed lens, and explore various financial, emotional, and social consequences to being moderated – including exacerbation of trauma and violence. Further, we draw on Gerrard and Thornham [39]'s ideas of 'sexist assemblages', reflecting on how content moderation processes reproduce larger systems of institutional power[20]. In other words, we consider how content moderation can be seen as the online manifestation of other forms of policing that impact oppressed groups in complex ways.

We also build upon best practices for co-creation and collaborative knowledge production that has a long history at CSCW [13, 29]. Part of the contribution of this work is a demonstration of how to apply Chen et al. [16]'s principles for trauma-informed computing when evaluating the effects of a technology for research. We draw upon this work through co-written case studies that specifically demonstrate harm caused by discriminatory content moderation, conducting research using the trauma-informed guidelines for qualitative research that Chen et al. [16] build off of [98]. For example, one of Chen et al. [16]'s six principles for trauma-informed computing is Collaboration, described as "ensuring that trauma survivors are actively involved in decisions regarding their care and support. In essence, trauma survivors should be treated as 'experts in their own lives'[75], which means recognizing and valuing their opinions and decisions." For computing specifically, this includes "ensuring survivors have representation and input during the development and evaluation of new technologies". We employ the principle of collaboration, along with participatory research and phenomenological methods, to arrive at the case studies presented in this work. Working with survivors of trauma, we also prioritized *consent-focused methods*, including content warnings, multiple opportunities for anonymity, full control over narrative presentation and the reporting of demographics information, and transparency through the process. We detail these methodological contributions in Section 3.

We deliberately chose to prioritize depth over breadth for this work, relying on relationship building with our participants to fully understand repercussions of discriminatory content moderation across multiple dimensions. Through particularizing these in-depth and specific narratives, we surface the details and *embodiment* of discriminatory content moderation on marginalized individuals. Through counter-storytelling [23, 25] and the situated knowledge [46] of our participants, we are able to gain insight into how discriminatory content moderation interacts with each of our

participant's unique relationships to systems of power. Through particularization and participatory research we are able to identify gaps in current scholarship on discriminatory content moderation. We also explore the depth of experience through centering our participants and their needs in a collaborative process. We provide rich detail of the point of contact between a social media platform with billions of users and those who experience marginalization across multiple planes. In doing so, we illuminate how deeply entangled 'online' and 'offline' harms are, arguing that they are not separate entities but rather enmeshed together in embodied experiences of discrimination. This perspective was possible through close, trusted and long-term relationships with our participants, who divulged the traumatic impacts of discriminatory content moderation over the course of two years.

The motivating question for this work is as follows:

> **RQ:** How do the lived and situated experiences of marginalized creators provide insight into the lasting impacts of discriminatory content moderation on user wellbeing?

The five narratives documented here provide jumping off points for future exploration of discriminatory content moderation, specifically its long-term impacts and enmeshment with trauma and users' relationships to systems of oppression. While the roots of discriminatory content moderation stem from larger dynamics of oppression [37], we highlight specific affordances of the IG platform that compound harm in these cases. Finally, we, alongside our participants, draw the narratives and larger context of work in this domain together to suggest future pathways for accountability, repair, and transformative justice. Throughout, we present avenues for participatory and collective advocacy that centers marginalized communities, trusting their expertise of discriminatory experiences on social media.

2 RELATED WORK

We are interested in the lived experiences and lasting impacts of discriminatory content moderation to creators with a variety of marginalized identities. We first review existing work that has illuminated the many ways that platforms employ automated and manual content moderation processes. This includes reviewing the history of content moderation practice, as well as considering the legal and social pressures for certain types of content to be moderated. Next, we cover scholarship that directly looks at how content moderation can perpetuate discrimination to marginalized groups. While prior work provides notable insight into both user experiences of moderation and the ways in which moderation can be discriminatory, we argue that a more in-depth look at these lasting impacts and how they are entangled with identity and power is beneficial for understanding this complex phenomena.

2.1 Discriminatory Content Moderation

Content moderation refers to "monitoring and vetting user-generated content for social media platforms of all types, in order to ensure that the content complies with legal and regulatory exigencies, site/community guidelines, user agreements, and falls within norms of taste and acceptability for that site and its cultural context" [78].Moderation actions include flagging hate speech, misinformation, nudity, cyberbullying, abuse, illegal activity, copyright infringement, and spam. Human moderators used to be solely responsible for content moderation on many online platforms, and often suffered psychological trauma from exposure to disturbing content [5, 9]. Today, many content moderation decisions are automated in some fashion; this switch to automation allows for increased scalability at a low cost [41]. Semi-automated content moderation is now the norm

on most large platforms; in this approach, a variety of algorithms are used to identify undesirable content which is then passed to human moderators to judge. Here we focus on Instagram, which has one set of Community Guidelines applying to all content on the platform, as opposed to other platforms like reddit that allow for tailored intracommunity moderation, which poses different challenges [36, 84].

At Instagram, moderation decisions are made with respect to the Community Guidelines provided in their Help Center; these guidelines are the main source of information for users regarding IG policies. The Community Guidelines prohibit: nudity, purchasing likes/followers, copyright infringement, selling drugs, supporting or organizing hate groups, offering sexual services, selling regulated goods, blackmailing others, threatening others with 'credible, serious threats of harm', glorifying self-injury, promoting eating disorders, or sharing videos or images of graphic violence. Any user on IG can report another user's content or account via a built-in reporting option. If a user is on the receiving end of a report, they may receive a Violation in their account, usually specifying which Guideline was breached. The user then has access to a Violations tab to keep track of which content was ruled as violating Community Guidelines. With enough Violations, the user will receive a message saying *'Your account may be at risk for deletion'*. Typically, users can appeal Violations via the Help Center, but research suggests this is often unreliable or may not even appear as an option for some users [45, 70, 94]. Content moderation is not always a strict removal of content – some content moderation may consist of 'flagging' the content with a banner, pop-up, or click-to-view functionality. On Instagram, a photo may be blurred and labeled as 'Sensitive Content' and the user can click to reveal the image. Other examples of flagged content are redirection to the CDC for COVID-19 related content or Twitter's flag for potentially misleading or false information.

Recent work has highlighted the ways in which content moderation is discriminatory, with disproportionate removals of Black users, Trans users, activists, LGBTQ speech, sex workers, sexual educators, and the infamous cases of moderated female body hair or period blood [3, 4, 11, 21, 27, 34, 45, 61]. There is evidence to suggest "double standards" in content moderation, with certain similar cases treated differently than others [28]. This discrimination is sometimes attributed to some or all of the following elements: algorithmic oversights, human moderator bias, and the ways in which social media users take advantage of the reporting features on the platform. Errors such as mistagging or punishing in-group members for language patterns are often ascribed to improper training data or an inability of algorithms to interpret context [55]. Audits for machine learning bias can help us understand likelihoods of discrimination in automatic content moderation (e.g. see [77]), though it is difficult to generalize to the real-world effects and frequencies [10]. Furthermore, borderline content being treated as a violation of guidelines may also be the result of normative cultural values reproduced by platforms in ways that are nuanced and difficult to simply categorize as an "error" [59]. For example, the conflation of educational sexual imagery with prohibited sexual content could easily be the result of sex negative values as opposed to a "bug" or "error". We see that discriminatory content moderation can happen due to technical, legal, political, or social reasons, and is further complicated by community reporting by users, with users engaging in 'report bombing' to target accounts and get them deleted [100]). It is also complicated by recent push to moderate misinformation, conspiracy theories, and rumors with particular regard to the information about the COVID-19 pandemic and COVID vaccine, as well as civic and electoral processes globally [7, 53, 67, 80, 93]. Research on discriminatory content moderation has focused on some of these various pressures, as well as the different types of content that gets moderated. Here we briefly summarize evidence of discriminatory content moderation on social media across various topics.

2.1.1 Moderating Hate Speech. Underlying many machine learning-based systems of automated content moderation are methods and tools for natural language processing (NLP), which can be used (with varying levels of success) to identify extremist speech, cyberbullying, or online harassment. For example, detecting hate speech and offensive language online in order to try to avoid abuse, racial slurs, and violence is an active area of work [60, 96]. Defining and bounding discriminatory content moderation is complicated by the complexities and challenges of attempting to moderate content at such a large scale. For example, hate speech moderation is used to remove text that *"encourages violence or attacks anyone based on race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases"* (IG Community Guidlines). To detect and remove hate speech one must first agree on how hate speech should be defined [58] – should the definition include humor, factual statements, rely on additional context, or specifically target protected groups? These are active research questions.

Despite the difficulty of agreeing on a definition of hate speech, practical action is taken via automated detection and removal of hate speech on social media, prompting legal and social debate [97]. This automation gives rise to technical issues such as sampling and annotating the proper training data, attending to context, and managing the trade-off between false negatives and false positives. While high accuracy has been shown to be achievable on current hate speech databases, our work touches on the impact of false positives on marginalized individuals. To illustrate, consider recent efforts by many communities to reclaim slurs as a form of empowerment, e.g. the Black community reclaiming variations of the n-word or the LGBTQ+ community reclaiming "queer" or "dyke". Mozafari et al. [68] and Dias Oliva et al. [27] found that automated hate speech detectors were highly biased against groups using reclaimed terms in a non-offensive way. Scheuerman et al. [81] also documents the various harms to trans individuals in technological spaces, noting how hate speech policies do not sufficiently engage with intersectionality.

Automated hate speech detection is limited, with several scholars questioning if it can ever be properly executed at scale [41, 71]. Olteanu et al. [74] describe how the mathematical costs may differ greatly from the perceived cost, harm, and impact to the user, and call for more human-centered risk assessment. Several scholars go further to imply that automated content moderation of hate speech and other content is actually exacerbating, rather than relieving, issues of content policy. This is due to the increased opacity of an already poorly understood process, as well as further complication of already existing inequities for marginalized users on social platforms [43].

2.1.2 Policing the Body. Social media platforms have a long history of policing images of the body, such as not allowing images of tampon strings or period blood, banning images of hair near the pubic area, and specifically banning "female nipples" [34]. Image classification is often used for the detection of nudity or 'sexual solicitation', though these are often conflated [4, 32, 34, 43]. In attempts to remove sexual content, platform moderation has been shown to disproportionately targets sex workers, sex educators, queer models, body positive accounts, and even sexual assault survivor accounts [3, 30]. Prior work showed how LGBTQ+ content creators are often automatically categorized as "adult content" on Youtube, even if they are children's educators or making videos about something benign. If they mention queerness, it can be marked 18+, even if their content is not sexual in nature [90].

Scholars have also investigated social media's role in providing information on sexual health as social media is often the main source of sexual health information for minors. According to Borrás Pérez [12], the censorship of sexual health education and sexually-related content on social media platforms disproportionately impacts LGBTQ+ youth. Policies surrounding nudity and sexual content tend to punish those who do not conform to normative ideals regarding sex, sexuality, and nudity, while reifying bodies that do, hindering any non-normative expression of sexuality

[39]. For example, Are [4]'s autoethnography work details how pole dancing is moderated on IG – threatening user rights, autonomy, and opportunity. Duguay et al. [32] investigates the platform values enforced via content moderation, community reporting, and content visibility, with a focus on queer women's experiences.

Black women in particular are subject to an overly disparaging gaze – with both automated systems and human moderators seeing them as more inherently sexual, threatening, and aggressive than white women [45, 61]. This perception of aggression is reflected in discriminatory moderation of what is considered *violent* – with anti-racism educators experiencing content removals when speaking out about the violence they themselves have experienced. Not only are racialized people at risk for being reported or harassed, they also conduct a lot of unpaid labor for the platforms – flagging racist comments and posts simply to keep their own community safer [85]. While marginalized creators experience death threats, rape threats, and other forms of abuse that may not be removed, their content can be taken down for anti-racist education [18, 63].

2.1.3 Violence and Harm. The moderation of violence and harm is a difficult task, especially considering the wide range of what can be considered violence [82]. Díaz and Hecht-Felella [28] demonstrate this complexity and the 'double standards' of platform governance with a case study on removals of potential ISIS involvement vs. white nationalists on Facebook, stating that "rules against terrorist and violent extremist content remain opaque, failing to provide clarity on which groups have been designated as terrorist organizations and granting the platforms immense discretion in enforcement". IG bans images and videos of graphic violence, but prior work has shown that videos of police brutality often go viral [22, 47] – a traumatizing experience for many viewers of that content, despite its potential for raising awareness.

There are also different forms of violence, such as self-harm or eating disordered behaviors. IG prohibits the glorification of self-harm, including eating disorders, stating in Community Guidelines: *"Encouraging or urging people to embrace self-injury is counter to this environment of support, and we'll remove it or disable accounts if it's reported to us"*. Moderation is also dependent on how one defines an eating disorder, and it has been demonstrated that content moderation may actually reassert that 'real' eating disorders only look a certain way [35]. We know that many people use social media to disclose and discuss depression and mental health concerns; they may be at risk of being reported or flagged, especially when their speech is more negative and less positively received by others [56].

2.1.4 Community Reporting. Social consensus and community reporting play a large role in what gets moderated, sometimes in discriminatory ways. IG urges in Community Guidelines: *"Help us keep the community strong: Each of us is an important part of the IG community. If you see something that you think may violate our guidelines, please help us by using our built-in reporting option."* However, this opens the door for reproducing the very systems of violence they aim to reduce on their platforms. For example, people may report images of fat creators as nudity or sexual content, whereas they wouldn't report a thinner model wearing the same kind of clothing [31]. Zeng and Kaye [100] detail how "report bombing" can be used to target creators with marginalized identities. Content being removed for seemingly no reason or an obviously inapplicable reason may have to do with how users organize and take advantage of the system to report accounts en masse, even if a violation has not occurred [18, 83]. These instances cannot be disentangled from systems of power and domination, as accounts owned by marginalized users are more likely to experience this type of targeted reporting than those who hold normative identities [6]. Paralleling themes presented subsequently, Zeng and Kaye [100] report how *"Interviewees were exasperated that reporting was effective at removing their own videos that did not violate guidelines but ineffective at removing content they reported for being harmful, problematic and legitimately in violation of community guidelines"*.

2.2 Beyond the Point of Moderation

Prior work has documented the various ways in which content moderation can disparately affect those with marginalized identities. However, we know less about the lasting impacts of moderation and how the experience of moderation interacts with marginalization in broader contexts. Myers West [70] investigates user experiences of account and content bans on social media, and pay careful attention to the consequences of content moderation beyond threats to freedom of speech. They state: *"Although many users did discuss their experience with content moderation as an inhibition of their capacity for self-expression, the accounts surfaced a wider spectrum of consequences, some of which were particularly detrimental to users who are already in a marginal position in society"*. In this work, the authors focus on the financial, affective, and functional consequences that occur from content moderation, both online and offline – though their work was not specifically on marginalized creators. Furthermore, they are limited by short user responses and survey data restricted only to the point of removals/moderation. Haimson et al. [45] demonstrates a convincing report on the frequencies of discriminatory content moderation for Black, transgender, and/or politically conservative users, and gives brief insight into the perceptions held by those being moderated. For example, they demonstrate how one Black participant describes the censorship on Facebook as "upholding of white supremacy and racism". One of the examples provided by Chen et al. [16] for further investigation is Content Moderation, where they state *"removal itself could be traumatizing. Removing content shared within these communities can also hamper peer support by decreasing the information and resources available, and decrease safety by severing vital connections with advocates and resources"*. This prediction will be evidenced in our results.

Building on these important prior studies, our work delves deeper into the entire narrative of how one experiences discriminatory content moderation – from the initial point of removal to how it interacts with these larger systems, perceptions, their communities, and one's own experiences of their trauma as affected by technology and beyond. Alongside our participants, we carefully map out how every point of moderation is also a reflection of larger systems of discrimination and power, and how moderation continues to affect their lives long after the initial removal.

3 METHODS

This study is designed to foreground the personal experiences of discriminatory content moderation, including the larger impacts that moderation has on marginalized individuals and their ability to do their work. Through interviews with marginalized creators we surface experiences and descriptions of each incident and provide knowledge that is embedded in that individual's identities, language, physical body, culture, and other experiences – an embodied and situated knowledge [46, 79]. In dialogue with our participants, we aim to gain deep understanding of their experiences through their stories, intonation, and gestures to illuminate the conditions of the phenomenon of discriminatory content moderation. Our research design draws on phenomenology, using disclosure of events as they appear to those who experience them [76]. Phenomenological studies aim to understand the subjective, lived experience of each of the study participants, and how the disclosure of these experiences can illuminate the conditions that allow for a particular phenomenon. We approach our participants' experiences with interpretive phenomenological analysis (IPA): "more likely to focus on how the whole experience is meaningful in the context of one's life as it has been, is being and might be lived." [86]

Phenomenological inquiry alongside our participants allow us as researchers to understand more about the situated and complex ways a phenomena like content moderation manifests for a marginalized individual [44, 66, 79, 91]. Participants are seen as co-researchers, and included in the crafting of narratives and analysis – with focus on the phenomenon as experienced by the

participants themselves. Our participants each have a variety of marginalized identities that are central to their work. Their experience of content moderation interacts with their experiences as marginalized individuals, and the traumas they have sustained [1].

We also draw from counter-storytelling methodology[8, 23, 25, 65], grounded in critical race theory [26, 62, 73]. Counter-storytelling is "used to magnify the stories, experiences, narratives, and truths of underprivileged communities" in opposition to dominant narratives, with dominant referring to "practices, norms, and ideas that have the most power and influence in social, institutional, and economic structures"[15]). In other words, counter-storytelling is "a method of telling the stories of those people whose experiences are not often told"[89], which includes raced, classed, and gendered individuals [64]. Counter-stories have been successfully employed in legal contexts to give insight into things like "disparate impact" [24]. One such example particularly relevant to CSCW is Ogbonnaya-Ogburu et al. [73]'s *Critical Race Theory for HCI*, which provides several narratives to highlight ongoing problems of race in HCI. Solorzano and Yosso [88] points to the multiple functions of counter-stories, including but not limited to: building community and solidarity among those at the margins of society, challenge dominant perceptions and transform established belief systems, and allow us to envision and build a richer reality that takes into account both the stories and our current systems.

In our work, each of the participants' experiences can be viewed through a lens of intersectionality. Intersectional analysis [17] resists looking at oppressive forces in isolation, and instead regards the interaction of multiple vectors of oppression and privilege to explore an individual's situated and subjective standpoint [20].

Our approach goes deep into how our participants experience their own content, moderation of that content, and the downstream effects of content moderation practices and policies. These counter-stories each provide a window into a richer reality of how these practices and policies perpetuate harm.

3.1 Participant Recruitment and Timeline

We recruited participants who led an anti-oppression advocacy or educational IG account and had publicly talked about an experience being banned or moderated on IG in the year prior. This moderation could apply to any user-generated-content, such as a post, story, or comment. One challenge inherent in working with vulnerable groups is to establish trust in the researcher-participant relationship. This is difficult without the researcher disclosing their motivations and prior experience. Therefore, the first two authors employed a recruitment strategy that was targeted; recruiting participants through direct invitation and word-of-mouth. They are creators that one of the authors had a prior relationship with, trusted their expertise, and knew that they had experienced discriminatory content moderation of various kinds. The two first authors are white, nonbinary-trans, and autistic, which was shared to participants following the feminist practice of reflevity and positionality statements. This undoubtedly had an effect on how participants chose to engage with us. Our research design, materials and methods explicitly addressed how we would be aware of, and counteract when possible, our own institutional privilege and the history of exploitative research on vulnerable populations [54].

We intentionally prioritize depth over breadth for this work. First, building trust and rapport takes intentional effort and time. Working with participants longitudinally allowed us to gain insight into the lasting effects of the content moderation experience. We explored how content moderation interacts with trauma [16]. Each of the narratives presented crafts a holistic view of an individual experience – something not captured by survey responses or frequencies of content removals. These narratives are not meant as a generalized view of discriminatory content moderation, though they do reflect some prior work's findings [45, 70]. Instead, they serve as a holistic and detailed view

of this phenomena, as told by users themselves. Each story contains details worthy of further study, potential interventions for repair, and future opportunities for community-based participatory research.

Our aim was to recruit and interview at least five participants. The degree – both frequency and duration – of engagement with each participant limits the scalability of this approach, making larger numbers of participants prohibitive. We have worked with these individuals for nearly two years (*21 months*), continually facilitating updates to their narratives, our conclusions, and how each individual is represented in the text. We continue to work with several of the participants, including updates of how their perspectives have changed over time. The initial updates to the narrative writing spanned several months of back and forth editing, followed by contact every few months over the past year. At times, the first two authors portrayed the participants with incorrect wording or assumptions. We were able to repair these mistakes due to the interpersonal relationships the first two authors had developed with each participant, and offered more opportunity for edits so that participants could be portrayed in their own words. The entire process consisted of the interviews, several narrative iterations, a follow-up survey, reporting of our findings to participants, iteration on those findings with participants, updates throughout the process of submission and review, and edits to how their identities and demographics were portrayed in the text – all with reassurance that participants could edit the text up until the final camera-ready version. In this way, our approach differed from standard interview practice. We expand on the reasoning for this in Section 3.3.

We conducted semi-structured interviews via Zoom, with the two lead authors and one participant per interview. Initial interviews lasted 60-75 minutes. Participants received $200 as compensation during the process. We took care to offer financial compensation reflective of the rates that some of our participants charge for consultancy, and to compensate for the time commitment involved in this project.

3.2 Participants

Our participants were recruited based on their expertise in activism work as well as work providing education and community spaces in the pursuit of dismantling systems of oppression. They each run accounts with large communities, and engage with their community through support groups, regular Live sessions, and other outreach. While providing participants demographics is common practice within research communities, we have found that this can feel reductive, often without consent of participants themselves. Demographic information is self-disclosed in the co-constructed narratives that follow. Here we provide an overview of the participants' work, as well as their Instagram accounts and their respective goals and audiences. It is important to note that we collaborated with participants for the following descriptions, with several iterations. In line with trauma-informed qualitative methodology [98], we collaborated diligently on how each participant is represented, translated, and shared in research texts.

Lauren is a certified victim advocate and trauma professional. Her account, MTMV Community Support Network, (mtmvcommunity) has 31K followers and centers the experiences of survivors of sexual violence through trauma-informed education and community peer support. Lilith is a model and activist, with her work centering the importance of representation and relationality. Her account, Lilith Fury | Goddess of Horror (lilith.fury) has 92.1K followers. She models for companies that do not have plus sized Indigenous and Latina representation. P3 holds a PhD in Black Studies and Women & Gender Studies. Her work focuses on the beauty and body politics of racism, and her account with upwards of 150K followers holds a mirror to white supremacy and whiteness. Constanza Eliana, an activist, writer, and educator, has 43.3K followers on the account Constanza Eliana | Decolonize (eliana.chinea). Through this account she works to decolonize wellness and

self-care practices as well as providing anti-racism education, specifically centering the nuanced racialized experiences of being a Puerto Rican non-Black POC in the United States. Tuck, a holistic sex educator, shares their journey with gender transition, sex, and relationships with their 9K followers on the account Tuck Malloy (intra_sensual), along with queer and trans inclusive sex education workshops, tutorials, and information.

3.3 Trauma-Informed Methods

Working with survivors of trauma and those with marginalized identities, we followed trauma-informed protocols for qualitative research [72, 98], as well as the six principles for a trauma-informed approach [72, 75]. This included acknowledging our institutional privilege at the start of each interview, and reiterating our commitments to participant *collaboration and safety* [16, 75]. We designed our interview process with warm-up and debrief as suggested by Nonomura et al. [72], and reminded participants that they could skip any question, end the interview and still receive compensation, or ask us any questions about the research. Our consent model was detailed and customizable, with one participant asking to be involved to a lesser degree. Phenomenology and counter-storytelling methods allow the participants to represent themselves in their own words, and this is particularly important for trauma survivors who have experienced mis-characterization, silencing, or vilification. Participants asked to be in contact with others in our communications, and after obtaining consent from each person we shared contact details among the group, in line with the principle of *peer support*. Practicing the trauma-informed principles of *empowerment & choice*, we offered multiple and ongoing opportunities to edit one's characterization, as well as multiple opportunities for anonymization. We found that through the process of this research, participants realized new experiences (such as how much they had dissociated from their fears of being moderated or the compounding effects of harm), anxieties around how they were represented for research, shifting goals for their pages and communities, and even new violations on their accounts. The research process can take a long time, and with continual updates with the participants we were able to see that content moderation was still an ongoing issue throughout, as well as still a source of stress and concern. For *transparency*, we made clear our limitations that we did not have the power to influence Instagram's decisions or restore accounts/posts, but that shedding light on these stories may be a step towards changes in policy and practice.

3.4 Interview Design and Execution

Interviews were divided into five sections of questions, each with a different goal and opportunity for relationship building with participants. **Introductions** gave the interviewers (the two lead authors) the opportunity to disclose their positionality. Interviewers reiterated informed consent before stating the research goals, emphasizing the participatory nature of the project. **Background** questions allowed the participants to share needed context about their online advocacy work. Participants were asked, for example, *"Tell us about your work on IG, including your audience, sponsorships, topics you cover, and the kinds of activities you engage in".*

The next section, **Content Moderation Experiences**, delved into the participants' encounters with content moderation. We defined content moderation for participants as: the process of determining whether user-generated content adheres to the platform's community guidelines (policies) by both humans and automated systems and removing content that goes against these policies. After reviewing terms (content moderation, ban, algorithm, shadowban), we asked questions including: *"What was the most notable time you had something deleted?"* Notable may refer to the most memorable, most emotional, most financially impactful, most egregious, or most publicly visible. We collected information about which content was deleted, which actions they have had blocked, and how it affected them. We did not specifically ask about or probe for discriminatory content

moderation experiences. The next section, **Theories About the Content Moderation Process**, investigated how our participants theorized about how content moderation works. However, in our interviews we learned that participants were not interested in the back-end mechanics of a harmful system, and did not want the responsibility of needing to know how it works. As interviews progressed, we eventually omitted most of this section, observing that attempting to involve our participants in the algorithmic underpinnings of the process came off as additional burden to them.

Finally, we concluded with a **Creative Cool-Down**. We acknowledged that the interview process can be charged and/or upsetting, and invited our participants to engage in some self-care questions and creative brainstorming. We asked about coping mechanisms for advocates facing content moderation. Next, we delved into the idea of repair and accountability from IG. We asked: "What actions, if any, could IG take to move towards repairing the harm they caused to you as well as to others more generally?". We posed a creative exercise: *"If you were to dream of your ultimate vision of an IG experience that is rewarding, validating, liberating, etc, what would that look like?"* It is important to us as researchers not only to uncover the experiences of harm, but to rely on participant narrative to collectively think about restorative design. We ended the interview by asking the participant if they could recommend anyone else for inclusion in the study.

3.5 Narrative Construction

The two lead authors separately listened to the audio recording of each interview, read the interview transcripts, and took reflexive notes – taking care not to analyze or generalize but to describe how the participant represented *themselves*. Interviews were automatically transcribed by Zoom software, but manually confirmed for correctness by the authors. In order to identify the main factors of each participants' story, the authors organized direct quotes and highlighted main themes and recurring descriptions of the experience. In writing the initial narratives, authors relied on both the interpreted main themes and direct quotes, in order to ensure that the narratives represented the data itself.

We engaged in a process of co-construction and verification for each narrative story by sharing writing with individual participants, listening to feedback, and revising. We compared to the original transcripts to ensure the result reflected the interview dialogue. Authors stayed in contact with each of the participants for over 12 months, and sent editable drafts of their story to each participant. Participants were asked to complete a feedback survey which included questions about if and how the participants would like to be anonymized in the story and resulting research outputs, and how the written narrative needed to be revised. Participants responded, for example, with quotes they wanted amplified and interpretations that did not resonate with them. Participants edited their own stories, and the authors updated the narratives where needed after confirming the changes did not contradict the original interview material. Finally, authors distilled the main theme of the study to emerge collectively across all of the stories: Discriminatory content moderation is deeply enmeshed with broader experiences of discrimination, and that online and offline harms cannot be separated. Rather, they should be treated as informing and affecting each other, with discriminatory content moderation having far-reaching and lasting impacts, including but not limited to trauma, financial loss, and significant behavior changes.

4 RESULTS

Below we share five personal narratives, co-constructed with each participant, and follow the inspired style of prior work Ogbonnaya-Ogburu et al. [73]. Each case study represents a counter-story [23, 25, 88], an account of the lived experience of being moderated online, and how that relates to race, gender, class, and other social categories of difference, alongside trauma. We rely on participants as experts of their own experiences [16, 75], and are concerned with their subjective

realities and the impacts of those perspectives. For each of the named authors represented in these narratives, we encourage readers to seek their content directly for the most accurate insight into their work, either on Instagram or through their respective organization home pages.

The following narratives contain mention of sexual violence, racial violence, fatphobia, transphobia, whorephobia, PTSD, police killings, and death threats.

4.1 Story 1: Silencing Survivors

Lauren is a victim advocate and certified trauma educator, though she points out that survivors do their own advocacy – she is there as a facilitator, an educator, an activist, and a community organizer. She runs the IG account @mtmvcommunity, formerly known as @metoomanyvoices, describing it as "an account that speaks to dismantling rape myths, consent education, sexual assault, and trauma education. Simply put it's a community support network for survivors and supporters of survivors." In the summer of 2021, Lauren went to access the account, as she did regularly, only to find that it had been deleted. IG said that the account "violated Community Guidelines" and that the "account had been deactivated". She recalls this moment and describes it: "I literally had just been on there, like before my eyes, I got banned and kicked off and it just didn't make any sense what was happening, I thought for sure that it was a mistake, it was not a mistake". The experience was retraumatizing. "It was truly horrific. I just didn't understand what was happening. It was scary and I felt a loss of control ... trauma is you know, the absence of choice and control being taken away from you, I felt like my heart broke." Her community, her voice, her sense of safety, was all taken away in a moment. The community, her community, was rocked – their stability disrupted. Many of them rely on @mtmvcommunity; checking in on the account is the "first thing [they do] in the morning" for peer support. Not only did Lauren lose access to the community she had inspired and cultivated, but she feared she was being targeted – potentially by her abuser.

Lauren describes having to carefully navigate 'trolls' and violence on her page. As a credentialed trauma educator she is equipped to intervene when teens troll her page with rape jokes – but she wondered if her willingness to intervene by calling schools and offering sexual violence prevention education was what led to the deletion of her account. "It was extremely violating, whoever it was; knowing that it's an attack on me and my community, people that I care about and survivors who are already violated enough. Just talking about it makes me nauseous how violating." She points to the real impact of social media and content moderation: "It's not a game, and I feel like a lot of people didn't understand how serious and how personal and scary it felt. That's my livelihood, that's my voice being threatened and taken away." She emphasizes the personal toll as well, describing how she "barely slept the five and a half days that my account was down in total, I barely ate, I barely slept. I was a mess."

In response, Lauren quickly did a "deep dive" on the internet. She reached out to lawyers, press, and created a second account to bring attention to the issue. She asked members of her community to send reports to IG stating that @mtmvcommunity had been taken down unfairly, using the 'Something's Not Working' option in the Help Center. Her community members also supported her using other strategies such as looking up who to tag, finding information on strategies, and reaching out to people they knew that might be able to help. Just as suddenly as she lost the account, it was reactivated. She has still not been told which Guideline was violated or what led to the reversal decision: "I still don't know. They as an entity never acknowledged that my account came back, was taken down; there was no follow up".

Lauren also emphasizes the impact of her community support and how important that is for trauma survivors, "There are many mitigating factors when it comes to the impact and symptomatology of trauma, one of them being how supportive and affirming those around you are in

the aftermath of traumatic experiences. It was a really traumatic experience that wrecked havoc on my nervous system at the time but the community came through for me in such an amazing way that I am not impacted by it anymore." She also highlights the juxtaposition between the impacts of this traumatic experience with with others, attributing her lack of long term effects to her community support. "I have thought about that is the big difference between my other past traumatic experiences that I am still working through is that I had people show up and care here."

She advocates for more direct access to communication with IG. "I understand that there's billions, literally billions, of users, but I think that there needs to be a way that people can actually speak to the company, the fact that there is not that's just a red flag." Expanding on this, she mentions how this lack of access to direct communication impacts her work, and for those who work with trauma survivors more broadly. "There's a lot of times where I'll have people that I have no idea who they are threatening to harm themselves. I shouldn't say a lot of time, but there have been instances where I'm truly afraid for other people's lives or I want to report them or try to get them more serious help, and IG takes zero responsibility for any of that." Lauren imagines some form of repair; she wants to "[get] answers" about what happened to her and why. "I don't even want an apology, because I don't think it would be sincere. I want acknowledgement that peoples' livelihoods, both financially and emotionally, are tied to these pages".

4.2 Story 2: Fat, Not Nude

Lilith runs an IG account, @lilith.fury, centered on plus-size representation, fat liberation, and demonstration of clothing and product options for disabled people. As a model and activist, she "tries to work with brands that either don't have plus size representation yet or don't have disabled people working with them or Latinas or Indigenous people". Coming from poverty, she doesn't try to influence people to spend money on something they can't afford, but rather wants to help people to see their options. The account is her main source of modeling income, though she shares stories of her life as a model, an actress, a mother, an autistic person, someone with lipedema, a STEM student, an advocate, a horror fan, a lesbian, and an online friend as well.

Lilith has had many IG accounts. In the beginning, she had a more personal account with pictures of her friends, her dog, and her at the beach. She describes how the personal "account got removed also for nudity which was really weird because it was just me and my dog – like mostly pictures of my dog and then there's a couple pictures where you could see my head. I didn't even want to show my neck because of really bad body dysmorphia". In another case, she had another account where she "started trying to get used to [her] body." "I'd have a couple of full body pictures and that one actually was a whole bunch of people just being racist and whatnot that's why I lost that account, they just spam reported me I guess."

As she became more comfortable with herself, she ventured into modeling and acting opportunities, but describes the need to be "under the radar" to keep her account. As a fat woman who "dares to exist" she is continually reported by both automatic content moderation and other users. Her accounts have been repeatedly deleted. She does not have access to Branded Content (a critical feature for her livelihood). She is paid less than more novice creators for videos with more views. Her body is scrutinized more violently even when she abides by the Community Guidelines. If she appeals reported violations, "it gets worse because not only do they ... remove that post anyway and tell me that I'm wrong, but then they're like well, while we're checking your account, we found a few more things so anytime you appeal it's like they retaliate against you, for daring to stand up against them". People flag her posts, referring to her weight as "self-harm"; they report her body as "nudity" even when she is fully covered.

Lilith is extremely careful to abide by Community Guidelines for fear of losing her account and her livelihood. She chooses the clothing she models, the brands she agrees to work with, and the

way she poses her body with policy guidelines in mind. Whether it's a pair of knee length shorts, a "hideous dress without cleavage", or a selfie in front of a nude statue at an art opening, her posts are frequently reported and deleted. She talks extensively about how she is not given equitable treatment under the Community Guidelines policy – stating that they "don't mean jack shit". She compares her experience to the many thin, conventionally attractive, white models who are never taken down for nudity, even when deliberately breaking the Guidelines. She is targeted, threatened, slandered, harassed, and abused online – all of these actions contributing to the deletion of her posts and risk to her livelihood. She says: "I don't want to lose my page. I'm terrified to lose my page. Every time IG goes down, I freak out, like oh my God, this is it, this is the day it's all gone. This is the day I lose my source of income, this is the day that I lose any and all opportunity, this is a day that everything is destroyed. And that is such a shitty feeling to constantly live in fear that your entire livelihood is not safe. It's not secure, and it can be taken away at any time. Just because somebody doesn't think that your body is worth being seen or that nothing you say is important because of how you look."

4.3 Story 3: Policed and Placated

P3 is a former therapist and cultural creative. She holds a PhD in Black Studies and Women & Gender Studies, and her work focuses on culture, identity, racial identity, as well as body and beauty politics (e.g. colorism, hair politics, skin bleaching). She is a "cultural critical source trying to dismantle white supremacy through teaching about white supremacy and our lived experiences". She talks about how white supremacy as a phrase can automatically put white people on the defense – "when they hear *white supremacy* they think Ku Klux Klan and they automatically say 'that's not me', but it *is* you. White supremacy does not just mean folks who rock the confederate flag or lynch people from trees. White supremacy is about how this is the lens through which you engage this entire world, and your entire existence." A key element of her story is highlighting the institutional racism embedded into the algorithmic systems as well as the decisions made by human beings within the platform, "The insulting piece is that you also have the audacity to come up with diversity statements, you also have the audacity to say black lives matter you also hear you know you have the audacity to talk about share black stories. You want a gold star for being a basically decent human being and you're not even there."

The story she shared with us began in the summer of 2020, which P3 refers to as the "Black Summer", with its increased white interest in Black pain. Protests against racism, and specifically anti-black police violence, broke out across the US, accompanied by sudden white urgency around racism, a more covert racist violence. Social media hashtags popped up overnight, including "amplify Black voices" and "share Black stories", P3 recalls. During this time, she was invited to a campaign where Black activists took over white influencers' IG handles as a means of amplifying the voices of Black women. What was overlooked by campaign organizers, was the fact that exposure can result in violence against people who hold marginalized identities, especially if they are outspoken about systems of oppression. P3 speaks to this fact in regards to her antiracism work on IG, "for a lot of folks they think exposure is something that we should all want and be excited about I am not that person primarily because of the work that I do, and I want to be able to talk and engage the way that I want to."

Prior to Black Summer, P3 had used her IG page to engage with her Black community – either through information about her ongoing work and projects or through relatable memes and celebration of Black. Following her participation in the campaign, her audience became murky and unwanted. P3 describes the experience as follows: "Prior to that summer... I had a better sense of who my community was. The exposure that came with the campaign brought lots of new followers. Gaining another 30 to 40,000 followers was not something that I celebrated, primarily because most

of them were not my target audience. My audience and all of the work that I do, for the most part, is Black people and/or folks who 'get it' whatever 'it' is." She describes the "critical and sometimes comedic eye with which I look through" and points out that the sudden interest in anti-racism that brought white people to her page also brought racist violence.

Following the campaign, she watched "in real time" as she was reported by a white woman for encouraging violence in a humorous post that was immediately taken down. She describes the post as "On surface what you saw was a Black man sitting in the front seat, looking at the camera being cute while two people are fighting in the backseat. It's fucking hilarious. Because most of us know that these two people are either brothers, cousins or friends" ... "But you, a white lady, come on and see two people fighting in the backseat of a car and asked me why am I 'condoning violence'." She exposes the reality of what moderating 'violence' on IG actually looks like, "Somebody walking around with an AK47 shooting somebody, is that violent? Absolutely yes. You allow the videos of a Black person being shot by the police, strangled, whatever, we get to watch those videos over and over and over again on your platform, but two boys fighting in a back seat is violence? Somebody in there knows which violence is okay according to their standards." In response to our description of 'unfair moderation' she stresses that this language is categorically incorrect: "I don't even know if it is a fairness involved – there's no *humanity.*"..." I don't have the word but fairness is too light."

In her telling of the experience, P3 draws attention to the entitlement that white people feel to police Black people, "I was very vocal on my page about [white supremacy and racism] as well. That makes white people uncomfortable. You want me to be anti racist in the ways you think is appropriate and that's not how it goes." Specifically, the report feature is available for everyone on IG but she highlights how white people have access to use it as a weapon similarly to how white people use the police to inflict racist violence on Black people, "No, this is not the white lady calling the cops on a black woman out in the world, but she called the cops on me [on IG]. In this day, it is the same shit and it's not just the calling of the cops. It's the fact that you believe you have the *right* to police me. You believe that whatever you say and see is wrong, is wrong. And guess what, there are people who will look at your white body and believe you before they asked me anything." "It's about the power and the privilege that is afforded whiteness irrespective of who you are. You have a voice. [IG] makes that very clear."

P3 also offers a broader critique of the platform,"[IG has] the capacity to make every update known to man on this app and you're telling me you don't have the capacity to engage this algorithm? It's because you don't want to – somehow it's *working* for you – say that." P3 highlights how these outcomes reflect the differential access to humanity for users, "For an algorithm, your very definition of a human being is white. Everyone else is some level of diversity". She would rather them call the Community Guidelines what they are – rules. She draws attention to the fundamentals of community that are missing, "Don't fake like we're community. We're not. Because if we were you would engage with the community differently. If these things are in place to somehow protect us, or to show us that you care, then why can't I engage with you?"

4.4 Story 4: Don't Say 'Decolonize'

Constanza Eliana's experience of content moderation sheds light on the weaknesses of both automatic content moderation and community policing. She explores the concepts of *fragility* and how much anti-racism educators are expected to censor themselves to be palatable to white audiences. Her initial goals around health and wellness were to increase representation for people of color in predominantly white wellness spaces – her work on the account @eliana.chinea quickly evolved into anti-racist and anti-colonial education. As a Puerto Rican she directly experiences colonization from the US Government, and trying to share her experiences within the predominantly white wellness industry was not welcomed: "you can't bring social justice or politics into wellness;

you shouldn't talk about it; you're going to lose students." Constanza Eliana and others argue that part of self-care, wellness, and healing is resisting colonization in all its various forms. This form of self-care is not separate from wellness, but rather an integral part of promoting well-being. In the beginning, she was careful to use "proper language in order to get people not to just see me as being aggressive or argumentative but to like really see that my experiences are validated by other people and validated by anti racist theory". Now, her anti-racist and anti-colonization education page addresses "the implications of colonization, and continued colonization, on different racialized identities, but also on different ethnicities and nationalities". Her content "is intended to educate the public around the experiences of marginalized identities and the experience of colonization."

As Constanza Eliana's content expanded to topics of racism and colonization, she started to get posts removed. The first was about Black Lives Matter, and "since then I've had at least 15 posts either deleted permanently or deleted temporarily." In describing the posts she has had taken down she says: "definitely everything has had some sort of racial component to it, calling out whiteness, calling out white supremacy." Constanza Eliana has a carefully curated community of educators, and has done work to hold people who are causing harm to her and her community accountable. One prevailing issue she highlights is that white people continue to profit and benefit off of anti-racist work – often in the place of a more qualified non-white educator. When Constanza Eliana drew attention to how white anti-racism accounts were misunderstanding concepts through their white lens and yet receiving book deals, she received major pushback. She was banned from using IG's Live feature for 2 weeks.

In sharing her emotional experience of content moderation Constanza Eliana discusses how it feels to be continually tone policed, gaslit, and silenced in the offline world. "Definitely the first reaction is anger because I'm really not doing anything wrong – because typically when you get banned or deleted or something like that it's because the platform thinks you've done something wrong, you violated something. And, for me, talking about social justice is not a violation, if anything, it should be talked about more. It also kind of messes with your value as well my sense of self worth because so much of what I do is tied up in not just the educational component, but my experience. So if something is being banned or taken down or shadowbanned as violating rules, meaning you've done something wrong. Then, that means my there's something wrong with my existence; my inherent experience. so definitely the mental health component of it is very tied to it as much as I do a lot of work around internalized oppression". These emotional responses dictate how much she is able to engage and continue her educational work. "It makes it too much of a burden and it's no longer you know fun for me to engage with my audience and engage with particularly other people of color and their experiences when I constantly have to worry about an algorithm you know either showing it or not, showing it banning it not banning it deleting it not deleting it like it's just too much of a burden."

Constanza Eliana has herself reported encountered instances of harm on IG, including a video of a white man in blackface and death threats she has received. These reports did not result in action by IG and she was told the content "did not go against Community Guidelines". She says "it's a very strange experience and so very because you know that social media isn't real and yet your real life has been threatened". She theorizes about why her content gets taken down and takes as an example posts that have included the term "white supremacy": "if I had to guess I would think that perhaps the platform, the content moderators, are within IG and Facebook and think that it's the use of those words that are causing the hatred or the violence or whatever, instead of actively taking a look at the people who really are white supremacists and over covert ways that are actually causing the violence". She goes on to point out that "so you can see, the power of white supremacy is coming into play: where they still decide what is racist and what is not; what is hate speech and what isn't." While she might be able to get around this censorship by purposefully

misspelling 'white', "Personally, I refuse to do that because I think it dilutes the essence of the education and what I'm trying to put out there: that 'white' is what needs to be censored, instead of *the action of whiteness* that needs to be repaired."Bringing her experience and expertise back to the Guidelines she says: "I think the way in which they are writing the guidelines gives too much room for the wrong people to be censored. So whoever is actually creating the rules, the guidelines, the policies isn't implementing anti racism into the the structure of the policies."

4.5 Story 5: Being Trans Isn't 'Bullying'

Tuck (@intra_sensual) is a certified holistic sex educator for mostly a queer audience. They use IG as a form of building community, promoting their business, and creatively exploring different kinds of educational content. Tuck began doing sex education work while at university – hosting small workshops on topics of consent and gender expression. They received pushback from the campus administration and decided to move their advocacy work to IG, which later became a way for them to support their own education-based business after losing their job due to the COVID pandemic. Starting out, "A lot of it was really personal. Because I was also exploring my gender identity and was exploring queerness a lot, and so it kind of came from that route of I'm just going to share what I'm learning about and what I'm teaching about and hopefully some other people will resonate with this". As their account grew, "Slowly the broader realm of the Internet started trickling into my life. It was at the point where I hit like 1 or 2000 followers that things started to get more messy on IG. It didn't feel as safe and comfy as being like 'Oh, these are just my friends and and you know my other sex educator communities that are following me'. There was more transphobia and more people pushing back against things that I was saying. I started to notice myself getting more anxious about the things that I was posting or the things that I was engaging with".

Tuck's content can at times push the boundaries of what is allowed by the Community Guidelines, and they have had many posts removed from the platform. "there's just been so many times [getting posts deleted], I feel like I have lost track at this point. The ones that hurt the most are like – there was one where I was posting something just about being trans and just being in gender exploration. And I really had no idea why somebody probably reported that. There was no nudity, it was pretty PG. I think I was writing about maybe some affirming sex that I had had, but that wasn't like definitely wasn't the main focus. And that got taken down and I just was like wow this sucks because this doesn't seem that edgy to me." The Community Guidelines that get cited in their experience seem incoherent: "It's usually like hate speech, or harassment and bullying. Those are the two ones that I often get for why this is being taken down. And that makes no sense. Whenever I see that I'm like well, but this is just such an incoherent definition of what hate speech is – clearly IG does not have an awareness of actual harassment and bullying."

In response, and as a way to manage the emotional impact, Tuck described dissociating from the constant battle with IG's content moderation. They feel "desensitized" – "a huge aspect of me staying on IG and maintaining my mental health is very much blocking some of these things out". They point out how "the systems of oppression and power stay in power because we're just trying to survive. So we're dissociating from these things and just trying to keep our heads down, and I think it can be really important to bring those feelings up, because then you get to engage with your rage."

For Tuck, "so many things come down to trying to eliminate sex from so many platforms." They point out the importance of examining who is making these decisions, and the power they have to do so: "I don't even know what's going on at IG behind the scenes at all ... I would also be really curious to know more about what the demographics are of that team; how they're making these choices; who they're getting money from." Tuck's repeated experiences with content moderation

inform ideas around accountability and repair, and include a transformative version of IG that is more consent-based, drawing on ideas of consent and care from sex. IG has an opportunity to "shift their whole platform setup to be more consent based rather than more censorship based ... I definitely could see it being valuable to have more kind of like consent based check ins with people as they're engaging with content. If you wanted to follow a page that had sexual content, having to maybe read a disclaimer – like this is going to engage with sexual content, you should be aware of it before you follow this person, if that is okay with you proceed." In imagining consent-centered design Tuck draws on and builds from their background in sex education and sexual safety, advocating for how repair might be achieved by embracing more consent-based values as opposed to censorship and punishment.

5 DISCUSSION

The above narratives cover a variety of experiences with content moderation – each giving insight into the lived realities of this particular form of institutional harm on the platform. These narratives uncover the complex and varied details of being moderated for one's body, race, trauma, or advocacy. We identify several dimensions of harm across these stories – financial impact, isolation from community, exacerbated trauma, privacy concerns, unpaid labor, and impacts to self-worth and self presentation. Participants tell us about losing access to sponsorships, unpaid labor of handling 'trolls', exacerbated trauma in the face of uncertainty and opacity, as well as the tradeoff between vulnerability and visibility. However, each individual's experience varies depending on their exact circumstance, relationships to power, and for what they were moderated. In presenting such different accounts, the variability of harm, the complexity and nuance of solutions and recommendations, and the larger systems of oppression at play are evident. In reflecting of these narratives we aim to move beyond simply reporting on discriminatory content moderation to look at the phenomena as a whole, focusing our lens on how it interacts with larger concepts of identity, privilege, and oppression. Each narrative reveals how content moderation reproduces systems of power and domination, blurring the arguably constructed boundary between the digital world and the "real world".

5.1 Online and Offline Harm Cannot Be Separated

For Lauren (Story 1), we saw how the experience of losing her account specifically triggered traumas around being silenced, targeted, and lacking control. She described the inability to eat and sleep during her attempts to learn and try everything she could in order to get her account back. She also demonstrated persistence and self-reliance, describing herself as a "fighter", because that is how she has had to take care of herself in the past. The fact that she was given no explanation for her account removal mirrored her lack of agency and voice as a survivor. Being cut off from her online community reflects the isolation and vulnerability she experienced after sharing her story. Lauren's experience of content moderation was directly linked to her experiences as a survivor of sexual violence.

Lilith (Story 2) also demonstrated persistence and unrelenting self-advocacy in the face of repeated discriminatory content moderation. This reflects the necessity for her to self-advocate to get proper medical care for her disabilities. Accustomed to not being listened to, she has developed strategies to continually fight for herself despite abuse and discrimination. This pattern of survival mirrors how she continued to find ways to support herself despite numerous account bans. It is also clear that Lilith experienced lack of access to means of success (that others had). On IG, she did not have access to branded content or the same payment rate as other creators. As a disabled Indigenous woman she has faced offline lack of access to resources, and she points out lack of equal opportunity online. Finally, her body is scrutinized more heavily than thin, white, able bodies –

both by algorithms and humans. This is true in the context of content moderation, given what is considered lewd or inappropriate, and also offline in the world by judgments and shaming of her appearance.

P3 (Story 3) points out how online reporting is another form of policing of Black communities. Black people are deemed more violent, threatening, aggressive, and dangerous – both online and off. While institutions may promote Black Lives Matter or Celebrating Black Voices, they still reproduce white supremacy through content moderation and other platform features. P3 would rather IG admit that their current algorithms and processes are *working* for them, than hide behind a placating facade. P3 had adapted a strategy of simply posting joyful memes and videos, including the video for which she was reported. This mirrors the expertise in the Black community and Black scholarship, which also has promoted Black joy as resistance and intracommunity solidarity in the face of oppressive online experiences [51, 69]. To have this joke used against her demonstrates that even Black joy can be seen as threatening. P3 chose to disengage from IG, demonstrating that in some cases institutional harm ruptures trust so severely that there can be no repair.

Constanza Eliana (Story 4) applies analysis of power to her experiences; as a non-Black woman of color, she is deeply in touch with the nuance of her identities. She describes how it is a "certain type of person who targets me", because most white supremacists tend to subscribe to a Black/white binary. She is able to view herself and her experiences through nuanced questions of privilege, identity, impacts, and politics. Her expertise in mental health and wellness also informs her strategies for coping with violence, while also illuminating how wellness spaces perpetuate racism. Constanza Eliana analyzes the technologies she engages with in a similar way to how she dissects racist politics. She identifies the avenues that allow for oppression and silencing, as well as the avenues that amplify more privileged voices. For example, she is able to look at how banning specific words may be helpful to avoid threats, but also circumvents accountability. In describing algorithmic precarity, she alludes to the volatile nature of public political opinion. She describes how one day her educational posts can be banned, while another day might receive 10,000 likes if it is a trendy topic. This mirrors the damaging trendiness of anti-racism when it is convenient for white people. Her analysis of IG's systems of content moderation were rich with insight into US politics, white fragility, and the nuanced experiences of being a Latina anti-racism activist.

Tuck's (Story 5) experience illuminates the conflation of gender, sex, and perceived threat. As a queer & trans sex educator, their content can be perceived as aggressive, inappropriate, and dangerous. Tuck was more comfortable in accepting that some of their sexual content might be received poorly by more conservative viewers, despite understanding that sex is not threatening or inappropriate. It was being moderated, reported, or harassed for their gender expression that was particularly painful. Offline, trans people are often viewed as inappropriate, sexual, dangerous to children and society, and not allowed to exist in certain spaces [79]. The policing of trans people may even be presented as "safety concerns", despite a lack of evidence to support these claims [2, 79, 87]. Anti-trans legislation tends to rely on laws regarding this safety, such as with bathroom bills, similarly to how content moderation is painted as a neutral or positive protection. The underlying rhetoric of anti-trans violence is that non-normative gender expression can be categorized as sexual aggression, fraudulent, or dangerous. These assumptions often lead to high rates of violence against and even death, specifically of Black, trans women [33, 79, 87]. For Tuck, the most painful points of moderation were when their authentic self was perceived as dangerous, just as how trans people are often regarded offline.

Each of these stories demonstrates how the experience of discriminatory content moderation is not limited to the online event of removal, but is rather an embodied experience, interacting with one's own identities, vulnerabilities, privileges, and perspectives. The effects can be longlasting, internalized, and must be situated in the larger context of how marginalized communities experience

oppression. Each instance of moderation also affects how creators choose to engage with their audiences and communities – often with fear of moderation affecting their authentic expression.

5.2 Content Moderation Related Affordances and Shortcomings of Instagram Platform

While it is crucial to recognize that harms perpetuated on social media platforms are embedded in larger systems of oppression, we identify some key features of the IG platform that exacerbate these harms to marginalized communities. Here we address our guiding research question through the *specific* technical affordances and shortcomings that our participants shed light on.

Community Reporting, and its unclear relationship to algorithmic moderation, surfaced in many interviews. While it is beneficial for users to have some form of reporting power on social media, Community Reporting (and other like tools) are vulnerable to coordinated action and manipulation which can be used to collectively bully, harass, or silence marginalized individuals. The forced choice categories presented to users making a report can also contribute to discrimination. For example, Lilith explained how people would report her as 'Self Harm' for being fat. P3 describes how 'Violence' has such a broad meaning, ranging from graphic violent imagery to someone taking offense to a humorous post. Lilith described an experience of being moderated for "sale of illegal or regulated goods" when modeling knee-length shorts for a company. Tuck is consistently moderated for "nudity or sexual activity" as a sex educator. Queer expression may also be automatically age restricted, or hidden behind a click-to-view "Sensitive Content" blur filter. Sensitive Content control may be an admirable endeavor, promoting user agency and control over their own IG experience, allowing them to hide, for example, guns/ firearms, nudity, or violence. All accounts are placed at a "Standard" level of sensitive content control by default, and the user may select "Less" or "More" sensitive content than "Standard". However, what is considered sensitive content, and automatically hidden without creator or user knowledge, can greatly impact marginalized creators' work. Constanza Eliana criticized this change when it happened, educating her community on the possibilities of her educational content being automatically hidden without her consent. The interaction between violations accrued and advertising revenue is also unclear, with one's ability to use Sponsored content at risk. For creators who rely on sponsorships for income, losing access to this feature translates into loss of earnings. Violations accrued may also affect the likelihood of being moderated on each subsequent incident; Lilith recalls getting a Story taken down for posing next to a nude statue – something she doesn't think would have happened if she hadn't already been moderated so often. Even this perception affects her willingness to show up authentically online, fearing lasting impacts to her financial position and opportunity.

Another main theme across participants is the lack of communication from the platform when a creator experiences content moderation. Lauren experienced drastic moderation in the form of account deletion, without any notification or explanation. Others seconded this sentiment, noting the IG Help Center as unreliable and sometimes antagonistic. While a global team of reviewers implies diverse perspectives, it is important to grapple with the fact that this is often exploitative, low-paid labor [40]. Moderators may also lack appropriate local context to accurately judge content, and incorporate their own biases into final decisions.

In order to keep their own communities safe, marginalized users engage in unpaid labor moderating their own pages. It is important to note that our participants often used the same features to do this, but had either opposing or incommensurable opinions about the utility of the features, highlighting that a "one size fits all" approach is not feasible. For example, our participants describe the complexity of being able to hide words in their comments, a feature intended for anti-bullying; Lilith hides threats of sexual violence, while Lauren may sometimes engage with threatening commenters as a form of public education. Constanza Eliana notes that some pages may hide words that hold them accountable, such as 'racist' or 'colonizer', restricting public accountability. Tuck

relies on quick Blocking when they encounter harassment. Restricting words becomes especially difficult when the words are used differently between in-group and out-group. For example, many queer people reclaim terms like dyke, fag, or slut.

Content moderation of billions of diverse users is a difficult task. It is needed to protect from egregious harm, and to ensure legality and safety of users. However, particular affordances of the platform stand out in the stories above. We also recognize that there was disagreement among our participants who each have their own identities, experiences, and values. This work is a starting point and motivates continuing this line of inquiry without treating marginalized creators as a monolith, and also providing spaces and resources for collaborative advocacy and resistance [95].

5.3 Limitations

This research is limited in the work it can do with and for vulnerable groups from the positionality of academia [54]. As we understand that academia and research and publishing practices are all institutions and processes that reproduce systems of power and domination, they are limited in their ability to work towards liberatory goals. Furthermore, both lead authors are white PhD students at a predominantly white institution. We acknowledge the importance of member research for building necessary trust for robust knowledge generation through interview studies. In one group session, it was noted by a participant that they felt uncomfortable being with a group of white people as a person of color. This arguably influenced their participation in the activities and therefore the data we obtained from them. While we co-authored with participants themselves, and shared other identities in common with participants, it is undeniable that future work should prioritize researchers of color (as funded researchers or paid consultants on the work). This study received limited funds and we chose to allocate them towards our participants.

6 POTENTIAL PATHWAYS FOR REPAIR ACCOUNTABILITY, AND TRANSFORMATION

6.1 Participant Recommendations for Repair

We were also interested in potential solutions coming directly from those experiencing the harm of content moderation. It may be the case that for some creators, the harm they experienced is beyond repair. This is because online policing is a larger reflection of systemic oppression far beyond social media, and trust has been repeatedly broken. Among our participants, the most common suggestion for accountability is for IG to interrogate their own practices, representation, and Community Guidelines. This is in line with Gerrard [38]'s six opportunities for feminist intervention for content moderation. Participants stressed that they wanted Instagram to acknowledge that content moderation impacts creators' livelihoods – both emotionally and financially.

Second, several of our participants requested transparency to users – explaining *how* their content gets flagged and *for what reason*, as well as *who* is making the policies or the final judgments. While we are seeing increasing pressure for platforms to increase transparency to researchers through data access, this same access is not applied to users. Suzor et al. [92] provides a framework for meaningful transparency, with recommendations for platforms to cite specific rules and reasons for violations, and Calleberg [14] stresses how users are much less likely to trust decisions from an AI. While scholarship continues to engage with these problems, there is still significant effort needed for progress to become apparent to vulnerable stakeholders.

Participants also mentioned the 'uselessness' of the Help Center, with no standardized turnaround or specific communication. Several participants wanted to "speak to a live human". Lilith urged for standardized payment for creators, after learning that another creator with less views and followers was receiving more money for a video. She also wanted the built-in ability to provide *evidence* for her case when she submitted appeals. Constanza Eliana critiqued the reporting categories,

demonstrating the wide range of how 'Violence' or 'Hate Speech' could be perceived or weaponized. It is unclear if there are specific teams who respond to types of reports. Tuck explored the idea of a more consent-based model for viewing content [48] – where the user is informed the kinds of sexual content that may appear from following a creator, and specifically opts in, perhaps even deciding how much they'd like to see.

Several participants commented on the cultural shifts required for any meaningful change – indicating how Instagram's profit model must be working for them properly and that so much would need to change for them to truly be accountable. This would involve hiring more BIPOC, disabled, trans, and sex-positive people in positions of power, working alongside creators who have been harmed, as well as applying a trauma-informed lens to the impacts of their decisions. Finally, accountability and repair may need to take the form of financial compensation – for the emotional and financial damages sustained.

It is crucial to view repair not just as immediate retribution, but also as a means of establishing safety, resolving uncertainty, and as a way to emotionally validate the experiences of the person harmed [99].

6.2 Call to Action for Researchers

Research has the power to shape platform design, policies, and features. Researchers can also focus work and resources on supporting community resistance and overall wellbeing for creative laborers and advocates.First and foremost our work highlights how intertwined 'online' and 'offline' experiences of technology are today. We deconstruct the boundary between harms experienced online and offline reality – and instead insist that *how* someone experiences harm is enmeshed in their identity, history, and positionality. This means that studying discriminatory content moderation must also attend to lasting impacts to one's financial security, access to community, and sense of identity and authenticity. Therefore, any ideation around repair or policy must consider the holistic view of those being affected.

We encourage researchers to think carefully about content moderation recommendations and to consider potential downstream effects to vulnerable populations. For example, increased moderation of potential misinformation could also result in more account bans for activists, particularly members of marginalized groups. Considering the potential consequences of a recommendation, and the financial, emotional, social, and mental burdens that could result, is a critical component. Basing recommendations for technical solutions solely on knowledge about bad actors does and will continue to harm marginalized creators. We also highlight the need for work examining lasting effects of content moderation, both on the intended problematic issues and unintended consequences.

Future work should attend to the immediate harms and need for repair in marginalized communities led by the communities themselves. This involves trusting the expertise of the people being impacted, and centering their experiences. For example, interviews should be trauma-informed and well compensated. IRBs should allow for co-authorship, co-construction of narratives, and participant-defined terms of consent. We see great potential for research that directly serves those harmed by discriminatory content moderation – for example, publishing narratives (with consent and participation) in order to help others feel less alone. We also encourage collaboration with other fields of research that have contended with issues of power and violence for decades. This includes feminist theory, critical race theory, disability studies, and queer and trans studies.

7 CONCLUSION

This research investigates discriminatory content moderation beyond the initial point of content removal – and considers the larger impacts, messages, and consequences of moderating members

of marginalized communities on social media and beyond. Through participatory and trauma-informed methods with Instagram creators, we present five co-written case studies of the lasting impacts of content moderation on those with multiple marginalized identities. We explore the re-traumatization that can occur from being moderated for one's identities and experiences. The narratives span topics of sexual violence survivor advocacy, beauty and body politics, anti-racist education, queer & trans representation, and sexual health education. Each of these counter-stories [23] demonstrates the long-lasting downstream effects of content moderation policy and practice. Our findings show how deeply entwined content moderation is with one's identities, communities, and personal history. We provide possibilities for accountability and repair, as well as outline opportunities for future research. We encourage future work that carefully considers the embodied experiences of those facing discriminatory content moderation on social media.

REFERENCES

[1] Sara Ahmed. 2006. Orientations: Toward a queer phenomenology. *GLQ: A journal of Lesbian and Gay Studies* 12, 4 (2006), 543–574.

[2] W Carsten Andresen. 2022. Research Note: Comparing the Gay and Trans Panic Defenses. *Women & Criminal Justice* 32, 1-2 (2022), 219–241.

[3] Carolina Are. 2020. How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist media studies* 20, 5 (2020), 741–744.

[4] Carolina Are. 2021. The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies* (2021), 1–18.

[5] Andrew Arsht and Daniel Etcovitch. 2018. The human cost of online content moderation. *Harvard Law Review Online, Harvard University, Cambridge, MA, USA. Retrieved from https://jolt. law. harvard. edu/digest/the-human-cost-ofonline-content-moderation* (2018).

[6] Imran Awan. 2014. Islamophobia and Twitter: A typology of online hate against Muslims on social media. *Policy & Internet* 6, 2 (2014), 133–150.

[7] Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. 2022. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour* (2022), 1–9.

[8] Thomas E Barone. 1992. Beyond theory and method: A case of critical storytelling. *Theory into practice* 31, 2 (1992), 142–146.

[9] Roukaya Benjelloun and Yassine Otheman. 2020. Psychological distress in a social media content moderator: A case report. (2020).

[10] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics*. Springer, 405–415.

[11] Danielle Blunt and Ariel Wolf. 2020. Erased: The impact of FOSTA-SESTA and the removal of Backpage on sex workers. *Anti-trafficking review* 14 (2020), 117–121.

[12] Pepe Borrás Pérez. 2021. Facebook doesn't like sexual health or sexual pleasure: Big tech's ambiguous content moderation policies and their impact on the sexual and reproductive health of the youth. *International Journal of Sexual Health* 33, 4 (2021), 550–554.

[13] Stacy M Branham, Anja Thieme, Lisa P Nathan, Steve Harrison, Deborah Tatar, and Patrick Olivier. 2014. Co-creating & identity-making in CSCW: revisiting ethics in design research. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*. 305–308.

[14] Erik Calleberg. 2021. Making Content Moderation Less Frustrating: How Do Users Experience Explanatory Human and AI Moderation Messages.

[15] M Castelli. 2021. Introduction to critical race theory and counter-storytelling.

[16] Janet X Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In *CHI Conference on Human Factors in Computing Systems*. 1–20.

[17] Ann-Dorte Christensen and Sune Qvotrup Jensen. 2012. Doing intersectional analysis: Methodological implications for qualitative research. *NORA-Nordic Journal of Feminist and Gender Research* 20, 2 (2012), 109–125.

[18] Kirsti K Cole. 2015. "It's like she's eager to be verbally abused": Twitter, trolls, and (en) gendering disciplinary rhetoric. *Feminist Media Studies* 15, 2 (2015), 356–358.

[19] Combahee River Collective. 1983. The Combahee river collective statement. *Home girls: A Black feminist anthology* 1 (1983), 264–274.

[20] Patricia Hill Collins. 1997. Comment on Hekman's" Truth and method: Feminist standpoint theory revisited": Where's the power? *Signs: Journal of Women in Culture and Society* 22, 2 (1997), 375–381.
[21] Kelley Cotter. 2023. "Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society* 26, 6 (2023), 1226–1243.
[22] Munmun De Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. 2016. Social media participation in an activist movement for racial equality. In *Tenth International AAAI Conference on Web and Social Media*.
[23] Richard Delgado. 1989. Storytelling for oppositionists and others: A plea for narrative. *Michigan law review* 87, 8 (1989), 2411–2441.
[24] Richard Delgado. 1990. When a story is just a story: Does voice really matter? *Virginia Law Review* (1990), 95–111.
[25] Richard Delgado. 1993. On telling stories in school: A reply to Farber and Sherry. *Vand. L. Rev.* 46 (1993), 665.
[26] Richard Delgado and Jean Stefancic. 2023. *Critical race theory: An introduction*. Vol. 87. NyU press.
[27] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25, 2 (2021), 700–732.
[28] Ángel Díaz and Laura Hecht-Felella. 2021. Double standards in social media content moderation. *Brennan Center for Justice at New York University School of Law. https://www. brennancenter. org/our-work/research-reports/double-standards-socialmedia-content-moderation* (2021).
[29] Catherine D'Ignazio, Erhardt Graeff, Christina N Harrington, and Daniela K Rosner. 2020. Toward equitable participatory design: Data feminism for CSCW amidst multiple pandemics. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 437–445.
[30] Christina Dinar. 2021. *The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act.* Technical Report. Technical Report. Heinrich-Böll-Stiftung.
[31] Lydia Dishman. 2019. This is the impact of Instagram's accidental fat-phobic algorithm. https://www.fastcompany.com/90415917/this-is-the-impact-of-instagrams-accidental-fat-phobic-algorithm
[32] Stefanie Duguay, Jean Burgess, and Nicolas Suzor. 2020. Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence* 26, 2 (2020), 237–252.
[33] Lee Edelman. 2004. No future. In *No Future*. Duke University Press.
[34] Gretchen Faust. 2017. Hair, blood and the nipple. In *Digital Environments*. transcript-Verlag, 159–170.
[35] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. 2020. Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.
[36] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.
[37] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511.
[38] Ysabel Gerrard. 2020. Social media content moderation: Six opportunities for feminist intervention. *Feminist Media Studies* 20, 5 (2020), 748–751.
[39] Ysabel Gerrard and Helen Thornham. 2020. Content moderation: Social media's sexist assemblages. *new media & society* 22, 7 (2020), 1266–1286.
[40] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.
[41] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234.
[42] Evelyn Nakano Glenn. 1999. The social construction and institutionalization of gender and race: An integrative framework. *Revisioning gender* (1999), 3–43.
[43] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.
[44] Kristin K Gundersen and Kristen L Zaleski. 2021. Posting the story of your sexual assault online: A phenomenological study of the aftermath. *Feminist Media Studies* 21, 5 (2021), 840–852.
[45] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.
[46] Donna Haraway. 2020. Situated knowledges: The science question in feminism and the privilege of partial perspective. In *Feminist theory reader*. Routledge, 303–310.
[47] Deion S Hawkins. 2022. "After Philando, I had to take a sick day to recover": Psychological distress, trauma and police brutality in the Black community. *Health communication* 37, 9 (2022), 1113–1122.
[48] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S Ackerman, and Eric Gilbert. 2021. Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms. In *Proceedings of*

the 2021 CHI Conference on Human Factors in Computing Systems. 1–18.
[49] 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. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.
[50] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.
[51] Javon Johnson. 2015. Black joy in the time of Ferguson. *QED: A Journal in GLBTQ Worldmaking* 2, 2 (2015), 177–183.
[52] Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Isabella Garcia-Camargo, Emma S Spiro, and Kate Starbird. 2022. Repeat Spreaders and Election Delegitimization: A Comprehensive Dataset of Misinformation Tweets from the 2020 US Election. *Journal of Quantitative Description: Digital Media* 2 (2022).
[53] Kolina Koltai, Rachel E Moran, and Izzi Grasso. 2022. Addressing the root of vaccine hesitancy during the COVID-19 pandemic. *XRDS: Crossroads, The ACM Magazine for Students* 28, 2 (2022), 34–38.
[54] Calvin A Liang, Sean A Munson, and Julie A Kientz. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 2 (2021), 1–47.
[55] Emma J Llansó. 2020. No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society* 7, 1 (2020), 2053951720920686.
[56] Mufan Luo and Jeffrey T Hancock. 2020. Self-disclosure and social media: motivations, mechanisms and psychological well-being. *Current Opinion in Psychology* 31 (2020), 110–115.
[57] Renkai Ma and Yubo Kou. 2021. " How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
[58] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one* 14, 8 (2019), e0221152.
[59] João Carlos Magalhães and Christian Katzenbach. 2020. Coronavirus and the frailness of platform governance. *Internet Policy Review* 9 (2020).
[60] Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427* (2017).
[61] Brandeis Marshall. 2021. *Algorithmic misogynoir in content moderation practice*. Technical Report. Technical Report. Heinrich-Böll-Stiftung.
[62] Aja Y Martinez. 2014. A plea for critical race theory counterstory: Stock story versus counterstory dialogues concerning Alejandra's" fit" in the academy. *Composition Studies* (2014), 33–55.
[63] Brad McKenna and Hameed Chughtai. 2020. Resistance and sexuality in virtual worlds: An LGBT perspective. *Computers in Human Behavior* 105 (2020), 106199.
[64] Lisa R Merriweather Hunn, Talmadge C Guy, and Elaine Mangliitz. 2006. Who can speak for whom? Using counter-storytelling to challenge racial hegemony. (2006).
[65] Richard Miller, Katrina Liu, and Arnetha F Ball. 2020. Critical counter-narrative as transformative methodology for educational equity. *Review of Research in Education* 44, 1 (2020), 269–300.
[66] Ryan A Miller and Annemarie Vaccaro. 2016. Queer student leaders of color: Leadership as authentic, collaborative, culturally competent. *Journal of Student Affairs Research and Practice* 53, 1 (2016), 39–50.
[67] Tamar Mitts, Nilima Pisharody, and Jacob Shapiro. 2022. Removal of Anti-Vaccine Content Impacts Social Media Discourse. In *14th ACM Web Science Conference 2022*. 319–326.
[68] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one* 15, 8 (2020), e0237861.
[69] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. 2022. Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 240, 17 pages. https://doi.org/10.1145/3491102.3517608
[70] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
[71] Yifat Nahmias and Maayan Perel. 2021. The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations. *Harv. J. on Legis.* 58 (2021), 145.
[72] R Nonomura, C Giesbrecht, T Jivraj, A Lapp, K Bax, A Jenney, K Scott, A Straatman, and L Baker. 2020. Toward a trauma-and violence-informed research ethics module: Considerations and recommendations. *London, ON: Centre for Research & Education on Violence Against Women & Children, Western University* (2020).
[73] Ihudiya Finda Ogbonnaya-Ogburu, Angela DR Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for HCI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.

[74] Alexandra Olteanu, Kartik Talamadupula, and Kush R Varshney. 2017. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on web science conference*. 405–406.
[75] A Treatment Improvement Protocol. 2014. Trauma-informed care in behavioral health services. *Rockville, USA: Substance Abuse and Mental Health Services Administration* (2014).
[76] Sadruddin Bahadur Qutoshi. 2018. Phenomenology: A philosophy and method of inquiry. *Journal of Education and Educational Development* 5, 1 (2018), 215–222.
[77] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.
[78] Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.
[79] Gayle Salamon. 2018. The life and death of Latisha King. In *The Life and Death of Latisha King*. New York University Press.
[80] Daniel A Salmon, Matthew Z Dudley, Jason M Glanz, and Saad B Omer. 2015. Vaccine hesitancy: causes, consequences, and a call to action. *Vaccine* 33 (2015), D66–D71.
[81] 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, Article 155 (nov 2018), 27 pages. https://doi.org/10.1145/3274424
[82] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.
[83] Joseph Seering. 2020. Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020), 107.
[84] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media and Society* 21 (01 2019), 146144481882131. https://doi.org/10.1177/1461444818821316
[85] Eugenia Siapera. 2022. AI Content Moderation, Racism and (de) Coloniality. *International Journal of Bullying Prevention* 4, 1 (2022), 55–65.
[86] Jonathan A Smith. 2011. Evaluating the contribution of interpretative phenomenological analysis. *Health psychology review* 5, 1 (2011), 9–27.
[87] C Riley Snorton and Jin Haritaworn. 2013. Trans necropolitics: A transnational reflection on violence, death, and the trans of color afterlife. In *The Transgender Studies Reader Remix*. Routledge, 305–316.
[88] Daniel G Solorzano and Tara J Yosso. 2001. Critical race and LatCrit theory and method: Counter-storytelling. *International journal of qualitative studies in education* 14, 4 (2001), 471–495.
[89] Daniel G Solórzano and Tara J Yosso. 2002. Critical race methodology: Counter-storytelling as an analytical framework for education research. *Qualitative inquiry* 8, 1 (2002), 23–44.
[90] Clare Southerton, Daniel Marshall, Peter Aggleton, Mary Lou Rasmussen, and Rob Cover. 2021. Restricted modes: Social media, content classification and LGBTQ sexual citizenship. *New Media & Society* 23, 5 (2021), 920–938.
[91] Wiley William Stem. 2020. *A Phenomenological Study of the Effects of Social Media Use on Minority Stress and Self-concept in LGB College Students*. Ph. D. Dissertation. New Mexico State University.
[92] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication* 13 (2019), 18.
[93] Jacqueline Urakami, Yeongdae Kim, Hiroki Oura, and Katie Seaborn. 2022. Finding Strategies Against Misinformation in Social Media: A Qualitative Study. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.
[94] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. " At the End of the Day Facebook Does What ItWants" How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on human-computer interaction* 4, CSCW2 (2020), 1–22.
[95] Julia Velkova and Anne Kaun. 2021. Algorithmic resistance: media practices and the politics of repair. *Information, Communication & Society* 24, 4 (2021), 523–540.
[96] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*. 19–26.
[97] Richard Ashby Wilson and Molly K Land. 2020. Hate speech on social media: Content moderation in context. *Conn. L. Rev.* 52 (2020), 1029.
[98] Rebecca Wong. 2021. Guidelines to Incorporate Trauma-Informed Care Strategies in Qualitative Research. (2021).
[99] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for

Computing Machinery, New York, NY, USA, Article 146, 15 pages. https://doi.org/10.1145/3491102.3517614
[100] Jing Zeng and D Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet* 14, 1 (2022), 79–95.

Received January 2023; revised July 2023; accepted November 2023