Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas

OLIVER L. HAIMSON, University of Michigan, USA
DANIEL DELMONACO, University of Michigan, USA
PEIPEI NIE, University of Michigan, USA
ANDREA WEGNER, University of Michigan, USA

Social media sites use content moderation to attempt to cultivate safe spaces with accurate information for their users. However, content moderation decisions may not be applied equally for all types of users, and may lead to disproportionate censorship related to people's genders, races, or political orientations. We conducted a mixed methods study involving qualitative and quantitative analysis of survey data to understand which types of social media users have content and accounts removed more frequently than others, what types of content and accounts are removed, and how content removed may differ between groups. We found that three groups of social media users in our dataset experienced content and account removals more often than others: political conservatives, transgender people, and Black people. However, the types of content removed from each group varied substantially. Conservative participants' removed content included content that was offensive or allegedly so, misinformation, Covid-related, adult, or hate speech. Transgender participants' content was often removed as adult despite following site guidelines, critical of a dominant group (e.g., men, white people), or specifically related to transgender or queer issues. Black participants' removed content was frequently related to racial justice or racism. More broadly, conservative participants' removals often involved harmful content removed according to site guidelines to create safe spaces with accurate information, while transgender and Black participants' removals often involved content related to expressing their marginalized identities that was removed despite following site policies or fell into content moderation gray areas. We discuss potential ways forward to make content moderation more equitable for marginalized social media users, such as embracing and designing specifically for content moderation gray areas.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: content moderation, social media, marginalization, misinformation, hate speech, transgender, Black, race, gender, LGBTQ, conservative, political orientation

ACM Reference Format:
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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 466 (October 2021), 35 pages. https://doi.org/10.1145/3479610

Authors' addresses: Oliver L. Haimson, haimson@umich.edu, University of Michigan, USA; Daniel Delmonaco, delmonac@umich.edu, University of Michigan, USA; Peipei Nie, niep@umich.edu, University of Michigan, USA; Andrea Wegner, amweg@umich.edu, University of Michigan, USA.

2573-0142/2021/10-ART466 $15.00
https://doi.org/10.1145/3479610

1 INTRODUCTION

People often disagree about what types of content are acceptable and unacceptable to post on social media sites. While one person may be excited to post a painting of a nude figure on Instagram, someone else may be offended by that painting appearing in their Instagram feed. Thus, each platform has developed rules and guidelines about what types of content are acceptable and what types of content should be removed. Some decisions about what should and should not be posted are made at the platform level via top-down approaches, while other decisions are made within online communities by moderators of those spaces [11, 47].

Throughout the popular media, stories have emerged about certain groups of social media users who perceive that their content and accounts are removed more often than others' (e.g., [14, 48]). Whether this is conservatives accusing Facebook of censoring conservative speech [71], racial minorities accusing Facebook of disproportionate censorship [49, 65], or transgender [1] people accusing TikTok of censoring trans content [20], it is clear that many different types of people feel that social media content moderation disproportionately impacts them in comparison to others. Studies have examined trans people's difficult experiences using social media sites [51, 104], especially related to account removals [54]. A recent report found evidence that women of color experienced higher rates of censorship on social media than the general population [102]. On the other hand, several reports have debunked claims that conservative content is removed more than content from people with other political orientations [4, 80]. In this work, we contribute an empirical examination of which types of social media users experience content and account removals more often than others, and what types of content are removed from each of these groups. We address three research questions:

RQ1: Which groups of social media users have content and accounts removed more often than others, for reasons they disagree with?
RQ2: What types of content and accounts are removed?
RQ3: How might types of content and accounts removed differ between groups?

We conducted quantitative (n = 909) and qualitative (n = 207) analysis of survey data and found that three groups of social media users experience content and account removals more often than others in our sample: conservatives, trans people, and Black people. While we did not initially set out to study these particular populations, this paper specifically focuses on these three groups because in our analysis to answer RQ1 we found that each of them disproportionately experienced removals. However, the types of content that people in each of these groups described having removed varied substantially between groups. Conservatives in our sample were more likely to have content removed that was offensive/inappropriate or allegedly so, adult, misinformation, or hate speech. However, trans participants were more likely to have content removed that was classified as adult despite following site guidelines, insulted or criticized a dominant group, or was specifically related to being trans or queer. Finally, Black participants were more likely to have racial justice or racism-related content removed. Taken together, these results indicate that while each group experienced content and account removals at high rates, content removals experienced by conservatives seemed related to platform and moderator attempts to remove harmful content and cultivate safe online spaces with accurate information, while content removals experienced by marginalized users (e.g., trans and Black people) tended to be content related to expressing participants' personal identities.

[1] Transgender refers to people whose current gender is different than their gender assigned at birth, including non-binary trans people. We shorten transgender to "trans" for the remainder of this paper.

In this work, we are especially interested in what we conceptualize as content moderation false positives and gray areas – content and accounts that are removed from a social media site yet do not actually violate site policies, or fall into a gray area with respect to site policies and community norms. Because online harassment is a pervasive problem, many researchers have studied how social media platforms and communities currently moderate to remove bad actors, and how these processes can be improved (e.g., [16, 67, 95]). Yet little research has focused on the content and people who are filtered out of online life due to content moderation false positives and gray areas – a circumstance that we find is more likely to occur for those who are marginalized or systemically oppressed due to their gender or race, and that can make these people further vulnerable [24, 36, 77, 99, 125]. In this research, we provide empirical evidence that marginalized social media users (i.e., trans and Black people) are more likely to fall victim to content moderation false positives and gray areas. By *marginalized* individuals, we mean the traditional definition of marginalized as being excluded or treated as peripheral to society, but additionally in this work we use marginalized to describe people who are systemically oppressed due to aspects of their personal identity (e.g., gender, race), who we find face particular challenges with content and account removals due to their identities. While some conservatives also consider themselves to be marginalized and victims of content moderation false positives and gray areas, we take a social justice perspective and center the concerns of oppressed people. We close by discussing how embracing gray areas may be a way forward to increase equitable content moderation for marginalized individuals and communities, and describe four potential approaches for consideration.

2 RELATED WORK

We provide a brief overview of content moderation research as it relates to marginalized groups, and then focus on prior work related to the three groups that we found experience disproportionate levels of social media content removals: conservatives, trans people, and Black people.

2.1 A Brief Overview of Content Moderation Research and Marginalized Groups

Content moderation is a fundamental part of social media platforms, a "blunt instrument" that is intended to be invisible as it removes content from online spaces [42, 43]. However, moderation becomes more visible for some users who face frequent removals [43, 115]. For instance, in the context of marginalized people's online narratives [15, 33], removing content and accounts takes away valuable support resources for people who seek shared non-normative experiences online [33]. Removals based on user reporting and flagging can act to favor majority norms and against marginalized groups, and harm the latter disproportionately [19, 35, 102]. As Duguay et al. [25] suggest, "...those who are compelled to flag others' photos do so because they feel strongly about the content, usually because they are offended by its violation of their personal norms, which may be sexist or homophobic." Participants in Fan and Zhang's [32] digital jury study expressed a similar lack of confidence in the quality and neutrality of user input, which echoes Park et al.'s [94] finding that crowdsourced approaches to comment moderation may show "undesired popularity bias." Overall, content flagging as a gatekeeping practice can privilege normative identities and experiences and disparage marginalized users [99], and can evolve into a form of digital gentrification that exacerbates power disparities [33, 76].

Reports have found that marginalized people are more likely to have content and accounts removed [35, 102, 125], and that they were often given only vague explanation (e.g., violated terms of service) or no explanation at all. In response, Suzor et al. [116] called for platforms to issue transparency reports about how they enforce their terms of service, and the Electronic Frontier Foundation launched the "TOSsed Out" project to track the ways in which terms of service are unfairly enforced [35]. Lack of transparency around content removals adversely impacts

marginalized groups [92, 99]. For example, content moderation practices often posit particular body types and experiences as normal and others as marginal [2, 3, 33], which can pressure some marginalized groups to conform and compel others to resist [33]. When pressured with conformity, marginalized groups may conceal stigmatized experiences [33], which can lead to psychological consequences related to hiding stigma [93]. For sex workers and others who rely on social media platforms to conduct work, frequent content and account removals can have substantial negative economic impacts [6, 7]. Marginalized groups can develop knowledges and skills through oppositional behavior that challenges inequality [22, 37, 38, 44, 82, 109, 126], such as politically significant social media-driven movements [15]. However, resistance is itself a burden, in that it requires marginalized people to work harder than dominant groups to experience the same levels of access to online spaces [2, 33]. Further, due to lack of transparency [66], little accountability, and the systematic failures of appeals processes [85, 122], users do not have the necessary resources to push back on oppressive platform practices [33]. Together, these factors contribute to inequitable social media content moderation practices for marginalized groups. Suzor [115] described the current content moderation landscape as "lawless:" social media platforms have tremendous power to make moderation decisions behind closed doors without relying on a consistent and vetted set of laws and processes. "Lawless" content moderation harms marginalized individuals and communities [115]. While previous research (described here) examined marginalized people's experiences with social media content moderation, we extend this line of work by quantifying the extent to which moderation inequities impact marginalized groups, and the types of social media content commonly removed, to inform more equitable future content moderation.

2.2 Political Conservatives and Social Media Content Moderation

Since 2016, conservatives have accused social media platforms of intentionally censoring conservative political speech [73, 113, 123]. They often base their claims on isolated cases in which conservative accounts were suspended from a site, did not show in search results, or were not being promoted sufficiently by a site's recommendation features [88]. Recently, several unprecedented cases further inflamed their claims. In 2020, Reddit banned its largest pro-Trump community "r/The_Donald" [63] after first quarantining the community [57]. In January 2021, Twitter permanently suspended @realDonaldTrump account [60] and Facebook subsequently suspended Trump's accounts indefinitely [129] after attaching warning labels on violating posts [100]. On the surface it may seem that social media platforms are censoring conservative speech, but examining the events that precipitated these moves reveals them as responses to repeated rule violations. These dynamics align with Seering et al.'s [106] findings that in online communities, "virtually all rule changes were made in response to unexpected incidents either gradually over time or suddenly following a specific incident." Researchers have argued that platforms' responses to defiance of rules constitute reasonable attempts to forestall further violence, and that they are not examples of ideologically motivated censorship [4].

To date, studies have found no evidence to support the claims that social media platforms unfairly remove or reduce distribution of conservative speech [83, 88]. Rather, a series of studies concluded that depending on the metric, right-leaning Facebook pages outperformed left-leaning pages or performed similarly [45, 79, 80]. More likely, platforms enforce rules based on the nature of the content in question. For instance, a recent study found that right-wing Twitter users spread misinformation *en masse* touting hydroxychloroquine as a Covid-19 remedy, sometimes drowning out expert information to the contrary, while Twitter attempted to remove this false and dangerous content [5]. In an investigation of comment moderation on YouTube, Jiang et al. [68] found that higher levels of misinformation, hate speech, and extreme partisanship resulted in heavier comment moderation for right-leaning videos. In a follow-up study reasoning about political bias in social

media, Jiang et al. [69] again showed that the likelihood of comment moderation on YouTube was equal across left- and right-leaning videos. Similarly, an audit of Google's search algorithm revealed that the algorithm was not biased along political lines but instead emphasized authoritative sources [83]. Another line of research indicates that the political right generates more online falsehoods than the left [74, 127], which inevitably leads to more content removals. For example, Kornbluh et al. [74] found that all of the top manipulators and false content purveyors on Facebook were right-leaning, and Zannettou [127] found that 72% of tweets with warning labels were shared by Republicans, while only 11% were shared by Democrats. Together, these studies helped explain why some conservatives believe their content is censored on partisan grounds when, in fact, it is being removed or demoted because it violates platform rules [4].

In response to widespread content moderation on mainstream platforms, conservatives have attempted to flee to alternative platforms like Parler and, more recently, Gettr [89]. Both promoted themselves as upholding free speech via lax moderation policies, yet quickly lost appeal for many users due to the toxic environments and rampant misinformation that emerged in the absence of effective moderation [89].

2.3 Trans People and Social Media Content Moderation

Trans people rely on online spaces for community, finding resources, and expressing changing identities [50, 104]. However, their online participation is often hindered by aggressive content moderation policies [20, 23, 54]. For example, Tumblr's automated filtering tools often mistakenly flag and remove trans content [36], and Facebook often removes trans accounts as being in violation of its "real name" policy, which simultaneously enforces and prevents online authenticity for trans users [54]. Facebook's insistence on users presenting one single identity is problematic for many users with faceted identities [75], including trans people [54]. Additionally, trans users often have accounts removed from Tinder after being reported as "fake" by other users [59, 101]. Trans users have also reported having content disproportionately removed on Instagram, where moderation algorithms appear to flag queer content more than non-queer content [14], and TikTok, where some posts were removed for violating site guidelines yet others were removed for no discernible reason [20]. Removing people and their content from social media sites either mistakenly or based on shortsighted policies that exclude trans users can be dangerous and can cause real harm for trans people, as content and account removals restrict access to online networks they need for support [77].

Often, when social media sites remove trans content, it is because such content is perceived to be "adult" [53]. Content related to trans surgery, sex education, or reproductive health may contain nudity, but requires different moderation policies and procedures than pornography [14, 53]. Nudity in trans surgery contexts is explicitly allowed on many social media sites, including Facebook and Tumblr. Nevertheless, sites have increasingly considered "adult" content threatening, and this has disproportionately impacted queer and trans content [14]. Even sites that welcomed sexual content in the past have recently changed course, such as with Tumblr's 2018 adult content ban [9, 53, 117]. This policy change made Tumblr's content moderation visible, allowing users to notice and critique its policies, processes, and politics [117]. The ways social media platforms classify content for moderation decisions reinforce norms, obscure complex genders and sexualities, and attempt to construct a "good" LGBTQ+ user by removing adult parts of online presentation [110]. In this way, social media platforms' control of sexual online content limits trans people's online presentation [9]. Trans people need stable online platforms that embrace adult content, and some have argued that platforms' business model should explicitly include such assurances [9, 53].

Trans content is also removed from social media sites due to sites' inability to distinguish between self-referential slurs used by trans and other LGBQ people, and slurs used as hate speech against

these groups [77]. For example, sentiment analysis methods may classify words like "tranny," "bitch," and "queer" as toxic even though these are self-referential words used frequently in queer linguistics [46]. Thus, if content moderation systems use computational linguistic techniques, this will hinder trans and LGBQ people's free speech and ability to reclaim language that has been used against them [46].

In addition to experiencing content and account removals, trans people are especially vulnerable to online harassment [67, 104]. In some online spaces, mocking trans and non-binary gender identities is commonplace [81]. Research has found that trans people have unique orientations to how they would want social media sites to respond to online harassment: trans participants were more likely than others to want solutions that educated others about trans identities and less likely to want more exposure after the incident [105].

Disproportionate social media content moderation can lead trans people and communities to feel invisible [103]. Unfortunately, trans erasure is nothing new: Namaste argued in 2000 that trans people "are perpetually *erased* in the cultural and institutional world" [87]. Trans people are underrepresented in the mainstream media, and thus social media platforms can help fill the gap; yet when content is censored, this limits representation and trans people are hindered from finding support and community [20]. In a recent article by online newsletter Salty, a trans and disability advocate described their experience being shadowbanned [2] on Instagram: "Hashtags help me reach my trans and disabled communities. However, because I was shadowbanned, my content no longer shows under the hashtags I use... I'm becoming invisible. My opportunities have been halted" [103]. To reduce trans invisibility, in this work we illuminate trans people's experiences with social media content and account removals to provide empirical evidence that supports the accounts of disproportionate censorship described above.

2.4 Black People and Social Media Content Moderation

Many social media users have drawn attention to instances in which platforms seem to disproportionately remove content from racial and ethnic minorities [1, 48, 49, 108]. For example, activist and podcast host Carolyn Wysinger found that her critical response to a racist post was removed from Facebook [49]. Social media content moderation algorithms have difficulty differentiating hate speech from discussion about race and often penalizes the groups they are supposed to protect [102]. Wysinger's content was removed for hate speech, and she was told she would receive a more extreme ban if she attempted to post it again [49]. As another example, Black TikTok users claim that the platform frequently shadowbans them and supresses their content, particularly when they post about race, racism, or Black Lives Matter [40]. In 2019 Facebook researchers found that under their new proposed automated moderation system for Instagram, Black people were 50% more likely than white people to have their accounts automatically disabled [108]. Rather than correcting this problem, superiors halted research on racial bias in Instagram's automated account removal systems, and eventually implemented a similar set of rules that was untested for racial bias [108].

The U.S. has a long and problematic history of racism, and the internet often mirrors this racism, becoming an outlet for racist content rather than an escape from it [98]. Even though racial justice movements have recently gained momentum and visibility, racist activity on social media has proven to be lasting [86]. Black women are often silenced online via content moderation in what Marshall [78] called "algorithmic misogynoir," which describes algorithmic practices that target Black women in particular. Moderators are sometimes trained to consider a post's popularity before taking it down, so racist posts that get a lot of likes are less likely to be removed [98]. Like trans

[2]Shadowbanning occurs when a person's content or account visibility is greatly reduced, or "content is made invisible to other users without actually being removed entirely" [85].

Table 1. Relationships between research questions, data collection methods, and data analysis methods.

Research Question	Data Collection	Analysis Method
RQ1	Survey 1 (n = 909)	Phase 1: Regression Analysis (Section 3.3.1)
RQ2	Survey 2 (n = 207)	Phase 2: Qualitative Analysis (Section 3.3.2)
RQ3	Survey 2 (n = 207)	Phase 3: Regression Analysis (Section 3.3.3)

people, Black social media users who experienced harassment were found to be less likely than others to want more exposure after the incident [105]. Ignoring ethnic and racial inequities in content moderation will continue to perpetuate injustices and keep some people's voices from being heard. Facebook's hate speech algorithms have traditionally treated content the same whether it described "female drivers," "Black children," or "white men," which, because disparaging comments about Black people tend to have more dire implications than comments criticizing white men, had the unintended effect of protecting dominant groups from hate speech more than marginalized groups [1]. Training tools also associate words commonly used in African American Vernacular English [107], such as the term "imma," with hate speech and threats even though the word itself has a neutral meaning [24]. However, Facebook is rehauling their hate speech algorithms in attempts to reduce these racial biases [26].

Disproportionate content removals and sites' inability to differentiate when someone is critiquing racism vs. being racist have caused Black activists and Black social media users to outwit Facebook's algorithms and moderators by using slang, emojis, and hashtags [49]. Black users also turn to aliases and back-up accounts to avoid losing their content and access to their community, even creating a system so fellow activists know when another is suspended and what posts led to their punishment [49]. However, the fact that racial minorities must employ workarounds like these simply to be able to use social media sites signals that content moderation inequities must be addressed.

While the articles we cite here from popular press sources indicate a trend in which racial and ethnic minorities' content is moderated more harshly than other social media users', we found little empirical research documenting these trends. Our study helps to fill this gap by empirically examining social media content and account removals experienced by Black social media users.

3 METHODS

We designed two surveys to ask individuals about their social media content and/or account takedown experience(s). Our goals were to compare different groups of users (e.g., based on demographics and political orientation) to see which may experience content and account removals at higher rates than others (RQ1), what sorts of content may be involved (RQ2), and how this might differ between groups (RQ3). All aspects of this study were reviewed and deemed exempt from oversight by our university's Institutional Review Board (IRB). Table 1 shows the relationships between our research questions, data collection methods, and data analysis methods.

3.1 Positionality

Given this paper's focus on certain groups, we felt it necessary to provide some information about our own identities in relation to those we study. The authors are all politically liberal, which necessarily impacted our analysis of participants' data across the political spectrum. We align with approaches like feminist standpoint theory that center marginalized experiences and acknowledge, rather than ignore, the social contexts in which research is conducted [55, 56]. Thus, though we do not consider it possible to be truly objective, we tried to reduce bias in our analysis, such as by considering data independently of participants' political orientations. We point out our political

orientations and other identity facets as a reflexive approach involving increasing context and highlighting potential biases rather than pretending there are none. The research team includes members of the LGBTQ+ community and has particular insight into trans participants' experiences. The authors are white and Asian, and we acknowledge the absence of Black research team members as a limitation of this study given the focus on Black experiences.

3.2 Data Collection

3.2.1 Survey 1. Survey 1 was designed to answer RQ1, and to act as a screening mechanism for Survey 2.

Recruitment. All Survey 1 participants were recruited through panel survey company Prolific. Participants were eligible for the survey if they were over the age of 18 and lived in the United States. Because there has been much popular press attention around particular marginalized groups disproportionately experiencing content moderation [14, 20, 49, 65], we oversampled for racial and ethnic minorities (n = 307), trans and/or non-binary people (n = 200), and LGBTQ people (n = 200) by specifically targeting each of these groups in Prolific's recruitment system. Because these groups tend to be more liberal than the general population, we also specifically sampled for conservatives (n = 100) so that we would have a range of respondents across the political spectrum. In early quantitative analyses we noticed that conservatives seemed to experience disproportionate amounts of content removals, which also influenced our decision to increase our sample of conservative participants. Finally, we sampled an additional 100 respondents without any particular focus, to ensure we included people of many potential demographic and political orientation combinations. Each participant could fall into multiple categories targeted in our recruitment strategy, thus the numbers of participants in each category exceed the targets listed here. Table 2 provides descriptive statistics about our sample. For some of the groups we purposely oversampled – trans and Black participants – we did find that they experienced disproportionate content moderation. For other groups that we purposely oversampled – non-trans LGBQ people, non-Black people of color – we did not find that they experienced disproportionate content moderation.

Survey Instrument. Survey 1 asked participants two primary yes or no questions: "Within the last year, have you had content taken down from a social media site for reasons you disagreed with?" and "Within the last year, has your account been taken down from a social media site for reasons you disagreed with?" Additionally, we asked participants about their demographic information (gender, race/ethnicity, age, income level, education level) and their political orientation (ranging from very liberal to very conservative). For the full survey instrument, see Appendix A. Some survey questions were loosely based on the questions from the survey [91] used in Myers West's study [85]. Before deploying, we pilot tested and workshopped our survey with several colleagues to ensure that the questions were easily interpretable. On average, the survey took participants 1.98 minutes (sd = 1.56), and each participant was compensated $.50.

3.2.2 Survey 2. Survey 2 was designed to answer RQ2 and RQ3.

Recruitment. Participants were recruited through panel survey companies Qualtrics (n = 70), Prolific (n = 125), and social media sites such as Twitter, Facebook, and Reddit by posting in online communities relevant to marginalized populations and through our extended personal networks (n = 12). Screening methods for our non-Prolific samples did not include all demographic and political orientation data (due to how the data was collected, such as through Qualtrics' internal panel survey screening mechanism), so those participants were excluded from the Survey 1 analysis. Participants were eligible for the survey if they had experienced either a content or an account takedown in the past year, were over the age of 18, and lived in the United States. On Prolific, participants were invited to take Survey 2 if they met the recruitment criteria when they took Survey 1; thus, the

Table 2. Participant Demographics

	Survey 1 (n = 909)	**Survey 2** (n = 207)
Gender		
woman	461 (50.7%)	98 (47.3%)
man	339 (37.3%)	80 (38.6%)
nonbinary	105 (11.6%)	30 (14.5%)
additional gender	6 (0.7%)	0 (0.0%)
did not disclose gender	4 (0.4%)	1 (0.1%)
transgender (lower bound)	151 (16.6%)	66 (31.9%)
Race/Ethnicity		
White	535 (58.9%)	97 (46.9%)
Black or African American	250 (27.5%)	100 (48.3%)
Hispanic or Latino/a/e/x	117 (12.9%)	17 (8.2%)
Asian or Asian American	92 (10.1%)	14 (6.8%)
American Indian or Alaska Native	32 (3.5%)	8 (3.9%)
Middle Eastern	14 (1.5%)	2 (0.9%)
Native Hawaiian or Pacific Islander	9 (1.0%)	2 (0.9%)
additional race/ethnicity	21 (2.3%)	0 (0.0%)
Age		
18-24	343 (37.7%)	66 (31.9%)
25-34	326 (35.9%)	71 (34.3%)
35-44	161 (17.7%)	41 (19.8%)
45-54	48 (5.3%)	13 (6.5%)
55-64	18 (2.0%)	9 (4.3%)
65 or older	13 (1.4%)	7 (3.4%)
Political orientation		
very liberal	261 (28.7%)	46 (22.2%)
liberal	289 (31.8%)	54 (26.1%)
moderate	204 (22.4%)	37 (17.9%)
conservative	126 (13.9%)	37 (17.9%)
very conservative	29 (3.2%)	33 (15.9%)
Intersectional identities		
transgender and Black	19 (2.1%)	5 (2.4%)
transgender and (conservative or very conservative)	10 (1.1%)	5 (2.4%)
Black and (conservative or very conservative)	25 (2.8%)	20 (9.7%)
transgender and Black and (conservative or very conservative)	3 (0.3%)	1 (0.5%)

Participants could choose multiple gender and race/ethnicity options, so percentages add up to greater than 100%.

groups that we targeted (racial/ethnic minorities, trans and/or non-binary people, LGBTQ+ people, and people across the political spectrum) were also targeted here based on their responses to Survey 1. For the Qualtrics survey panel, Qualtrics distributed the survey and screened participants using the same set of questions we used in Survey 1. We asked Qualtrics to recruit demographics similar to a U.S. representative sample, but with racial and ethnic minorities oversampled and trans and/or non-binary people included. Participants recruited via social media were screened for eligibility on the first page of the survey using the same questions from Survey 1. In this group, LGBTQ+ people were oversampled due to the online communities where we posted recruitment materials. While our full Survey 2 dataset included 326 responses, for this paper, we only analyzed the 207 Survey 2 responses from participants who (based on our Phase 1 analysis (see section 3.3.1)) were in groups disproportionately more likely than others to experience content or account removals (conservative, trans, and/or Black participants).

Survey Instrument. Survey 2 included 35 questions: 14 open-ended questions and 11 multiple choice questions about people's content moderation experiences, and 10 demographic questions. The parts of the survey instrument that were analyzed for this paper are included in Appendix A. Some questions were adapted from OnlineCensorship.org's survey [91] used in Myers West's study [85]. The survey was pilot tested and workshopped with a group of our colleagues who provided feedback on the question wording and the survey structure. Because most questions were open ended, we were able to read through responses and provide quality checks to make sure responses were reliable. We started with a small pilot sample of participants (n = 20) and carefully read through all responses to gauge whether participants seemed to be interpreting questions correctly before deploying with a larger sample. We closely monitored all survey responses throughout data collection to ensure data quality, and removed all responses where participants did not answer the questions or where text appeared to be gibberish or computer-generated. On average, the survey took participants 7.13 minutes (sd = 7.82). Prolific participants were compensated at a rate at or above $12 per hour. Qualtrics participants were compensated by Qualtrics directly. Participants recruited via social media were entered into a drawing for a $50 gift card.

Survey questions were designed to examine participants' experiences with social media content and/or account takedowns and their perceptions of how and why it happened. We asked users what reasoning was provided by the platform for why their content or account was removed, as well as why they thought it happened. This enabled us to compare the formal reason the content or account was removed with the user's perceptions and understandings of the situation. We also included questions regarding the personal impact the takedown had on the participants. This allowed us to understand potential short-term and lasting effects takedowns had on participants.

3.3 Data Analysis

We used a mixed methods approach to answer our research questions. First, in Phase 1, we used regression analysis on Survey 1 data to understand which groups were more likely to experience content and account removals (RQ1). Then, the answers to RQ1 informed our selection of data to analyze in Phase 2 to understand which types of content and accounts were removed (RQ2), for which we conducted qualitative analysis of Survey 2's open-ended questions. Finally, to understand differences between groups (RQ3), in Phase 3 we used the results from our qualitative analysis to build regression models.

3.3.1 Phase 1: Regression Analysis. To determine which groups were more likely to experience content and account removals, we built two logistic regression models using data from Survey 1 (n = 909). In Model 1 (see Table 4) the outcome variable was a binary indicator of whether or not that participant had experienced a *content* removal in the last year, and in Model 2 the outcome variable was a binary indicator of whether or not that participant had experienced an *account* removal in the last year. In both models, the independent variables included gender, race, sexual orientation, education level, income level, age, and political orientation. Reference categories were man, white, age 18-24, and politically moderate. Multicollinearity was not present in either model.

3.3.2 Phase 2: Qualitative Analysis. Phase 1's regression models indicated that conservative, trans, and Black participants were more likely to experience content and account removals than those not in those groups. Thus, for our next stage of analysis, we analyzed all responses to Survey 2 where the participant was trans, Black, and/or conservative or very conservative (n = 207). Our unit of analysis was at the participant level; each participant's six (if they had experienced a content removal or an account removal but not both) or twelve (if they had experienced both a content and account removal) open-ended survey responses were considered together. The first author began by reading through the data and conducting a first round of open coding [114]. Throughout the

process of open coding the full dataset, the first author created a codebook and used axial coding [114] to organize codes into categories. Next, the first author and the second author, who had also read through the dataset, discussed and collaboratively revised the codebook and categories, settling on four categories: types of content removed, account-related issues, interpersonal issues, and perceptions. The first author and the second author then each conducted a second round of coding on the full dataset, noting whether each code did or did not occur for each participant. The coders are highly knowledgeable about social media site guidelines due to prior relevant research projects, and considered and consulted site policies frequently throughout the coding process. During analysis, to the extent possible with the limited information we had, we verified whether content participants described was aligned with site guidelines. Some of the social media site policies relevant to our qualitative codes are detailed in Appendix B for the three sites most commonly used by participants: Facebook, Twitter, and Instagram. We discussed and resolved all instances of disagreement. The resulting dataset consisted of 0/1 indicators of whether each code applied to each participant's data. The final codebook is included in Table 5.

3.3.3 Phase 3: Regression Analysis. Next, we used the results from Phase 2's analysis to understand which codes were more likely to occur for each of our three groups of interest: conservatives, trans people, and Black people. Table 6 shows the results of these analyses. We built three logistic regression models, in which the outcome variables were whether or not a participant was conservative (Model 3), trans (Model 4), or Black (Model 5). Independent variables were the codes identified in Phase 2. We used lasso regression to determine which independent variables to include in each model. There were quite a few codes that appeared only in one group and not for the others (e.g., perceived anti-conservative bias only occurred for conservative participants, while only trans participants posted trans content), so initially our models suffered from inflated coefficients related to complete or quasi-complete separation. Thus, we removed all codes that demonstrated complete or quasi-complete separation, and instead indicated those instances by shading them gray in Table 5. Phase 3 resulted in an understanding of which codes were statistically more likely to occur for participants who were conservative, trans, or Black, as compared to those who were not. These are presented in bold text with significance stars in Table 5.

3.4 A Note on Self-Reported Data

Because content moderation is generally invisible for researchers like us who do not have access to social media platforms' behind-the-scenes content moderation logs, we relied on surveys for data collection. However, this means that all of our data is participants' *self-reported* experiences with content moderation. We were not able to confirm participants' reports with evidence of actual content and account removals. Thus, in this paper when we describe content and account removals, this means *participants' reports* of such instances.

4 RESULTS

Next, we will describe our results, providing empirical evidence about which types of people are more likely to have content and accounts removed from social media platforms, what types of content are removed for each of these groups, what account-related issues they experience, and some of the perceptions each group holds about their content moderation experiences. To begin with some descriptive statistics, Table 3 shows what platforms participants' content and account removals occurred on. Facebook was the most prevalent, followed by Twitter, Instagram, YouTube, Tumblr, and Reddit. A majority of participants' content removals (68%) were perceived as being removed by the platform itself, while around 23% were perceived as being removed by another

Table 3. Platforms participants' content and accounts were removed from, and participant perceptions of who removed their content.

	Content removed *n* (%)	**Account removed** *n* (%)
total *n*	200	90
Facebook	94 (47.0%)	49 (54.4%)
Twitter	38 (19.0%)	17 (18.9%)
Instagram	37 (18.5%)	16 (17.8%)
YouTube	18 (9.0%)	6 (6.7%)
Tumblr	18 (9.0%)	2 (2.2%)
Reddit	15 (7.5%)	4 (4.4%)
TikTok	11 (5.5%)	3 (3.3%)
WhatsApp	8 (4.0%)	2 (2.2%)
Discord	6 (3.0%)	1 (1.1%)
Snapchat	5 (2.5%)	1 (1.1%)
LinkedIn	4 (2.0%)	0 (0.0%)
Pinterest	3 (1.5%)	3 (3.3%)
Quora	3 (1.5%)	1 (1.1%)
other platforms mentioned by one participant each: Adam4Adam, Artstation, Amino, Deviant Art, KakaoTalk, Nimses, MeetMe, Mocospace, TripAdvisor		
content removed by platform	136 (68.0%)	
content removed by social media user in a moderator/admin role	45 (22.5%)	
participant unsure who removed content	27 (13.5%)	

social media user in a moderator or admin role. In roughly 14% of responses, participants were unsure of who removed their content.

4.1 Whose Content and Accounts are Removed?

Table 4 presents logistic regression models showing which groups are more likely to have content or accounts removed from social media sites for reasons they disagreed with. Model 1 addresses *content removals*, and indicates that trans people are significantly more likely to experience content removals than non-trans people ($\beta = 1.17$, $p < 0.001$), and that political conservatives are significantly more likely than political moderates to experience content removals ($\beta = 0.57$, $p < 0.05$). Overall, 29% of Survey 1 participants had experienced content removals. As shown in Figure 1, 34% of those who considered themselves conservative had content removed as compared to 30% of those who stated they were very liberal, 30% of liberals, 25% of moderates, and 34% of those who stated they were very conservative (this group was not statistically different from moderates in Model 1 due to a smaller sample size). 46% of trans people in our sample had content removed as compared to 26% of cisgender people[3]. Model 1 and Figure 1 do not show statistically significant difference between participants of different races/ethnicities.

Model 2 (Table 4) addresses *account removals*, and shows that three groups are significantly more likely to experience account removals: trans people ($\beta = 1.39$, $p < 0.001$, as compared to non-trans people), political conservatives ($\beta = 1.08$, $p < 0.01$, as compared to political moderates), and Black people ($\beta = .92$, $p < 0.05$, as compared to white people). Additionally, those who are nonbinary are less likely to experience account removals than those who are not. Overall, 12% of

[3]We refer to participants as "non-trans" if they did not answer "Yes" to the survey question "Are you transgender?" We refer to participants as "cisgender" if they did not answer "Yes" to the survey question "Are you transgender?" and also did not select "Non-binary" for the survey question "What is your gender." Thus, some nonbinary participants were not categorized as trans *or* cisgender, aligning with current evolving discussions that increasingly resist a cisgender / transgender binary. We acknowledge that our methods may unfortunately categorize some people incorrectly. In future work, we plan to adjust how we ask about trans status to better reflect these complexities.

Table 4. Logistic regression models examining factors associated with having social media content (Model 1) and accounts (Model 2) removed for reasons the participant disagreed with.

	Content removed (Model 1)			**Account removed (Model 2)**		
Variable	Coefficient (SE)	Odds ratio	95% CI	Coefficient (SE)	Odds ratio	95% CI
intercept	-1.78*** (0.34)	0.17	(0.09, 0.33)	-3.05*** (0.50)	0.05	(0.02, 0.12)
man (reference category)	-	-	-	-	-	-
woman	0.19 (0.17)	1.21	(0.87, 1.69)	0.11 (0.24)	1.12	(0.71, 1.79)
nonbinary	-0.30 (0.29)	0.74	(0.41, 1.30)	-1.04* (0.50)	0.35	(0.12, 0.90)
transgender	**1.17*** (0.23)**	**3.22**	**(2.06, 5.07)**	**1.39*** (0.32)**	**4.01**	**(2.14, 7.55)**
white (reference category)	-	-	-	-	-	-
Black	0.36† (0.19)	1.43	(0.99, 2.06)	**0.92* (0.26)**	**2.51**	**(1.51, 4.17)**
Latino/a/e/x	0.31 (0.23)	1.36	(0.85, 2.15)	0.05 (0.38)	1.05	(0.48, 2.13)
Asian	-0.17 (0.28)	0.84	(0.48, 1.43)	0.23 (0.40)	1.26	(0.55, 2.65)
Native American	0.09 (0.40)	1.09	(0.47, 2.35)	-0.65 (0.76)	0.52	(0.08, 1.85)
additional race	0.13 (0.50)	1.14	(0.39, 2.94)	0.28 (0.77)	1.32	(0.20, 4.95)
LGBQ	0.16 (0.20)	1.17	(0.79, 1.74)	0.31 (0.29)	1.36	(0.76, 2.41)
education level	0.02 (0.07)	1.02	(0.89, 1.16)	0.10 (0.10)	1.01	(0.83, 1.23)
income level	0.05 (0.04)	1.06	(0.97, 1.15)	0.02 (0.07)	1.02	(0.90, 1.16)
age 18-24 (reference category)	-	-	-	-	-	-
age 25-34	0.12 (0.19)	1.13	(0.78, 1.63)	0.20 (0.28)	1.22	(0.71, 2.10)
age 35-44	0.11 (0.23)	1.12	(0.71, 1.74)	0.26 (0.32)	1.30	(0.68, 2.43)
age 45-54	-0.49 (0.41)	0.61	(0.26, 1.31)	-0.30 (0.58)	0.74	(0.20, 2.13)
age 55+	-0.06 (0.46)	.094	(0.36, 2.23)	-0.52 (0.79)	0.60	(0.09, 2.32)
very liberal	0.00 (0.25)	1.00	(0.62, 1.62)	-0.67† (0.39)	0.51	(0.24, 1.09)
liberal	0.12 (0.22)	1.12	(0.73, 1.73)	0.12 (0.31)	1.13	(0.62, 2.10)
moderate (reference category)	-	-	-	-	-	-
conservative	**0.57* (0.27)**	**1.77**	**(1.04, 3.01)**	**1.08** (0.35)**	**2.94**	**(1.48, 5.94)**
very conservative	0.58 (0.44)	1.78	(0.73, 4.16)	0.60 (0.61)	1.83	(0.48, 5.63)
AIC	1100.20			633.58		
n	909			909		

† p <0.10; * p < .05; ** p < .01; *** p < .001

Notes: All variables were binary dummy variables except for two ordinal variables: education and income, which both ranged from 1-7. The transgender variable was not in comparison to the reference category. This was instead a binary measure of whether the participant stated they were transgender, regardless of whether they were a man, woman, and/or non-binary person.

Survey 1 participants had experienced account removals. As shown in Figure 1, 21% of those who considered themselves conservative had accounts removed as compared to 7% of those who stated they were very liberal, 13% of liberals, 10% of moderates, and 14% of those who stated they were very conservative. 19% of trans people in our sample had accounts removed as compared to 10% of cisgender people. 16% of Black participants had accounts removed as compared to 10% of white participants.

In this paper we chose to focus on political conservatives, trans people, and Black people because these results were statistically significant in Models 1 and/or 2 (Table 4). Although the statistical effects related to race are not as highly statistically significant as some of the other effects, we felt it was critical to focus on Black participants' experiences with content and account removals because addressing racial injustice and examining racial disparities in sociotechnical systems is important [90].

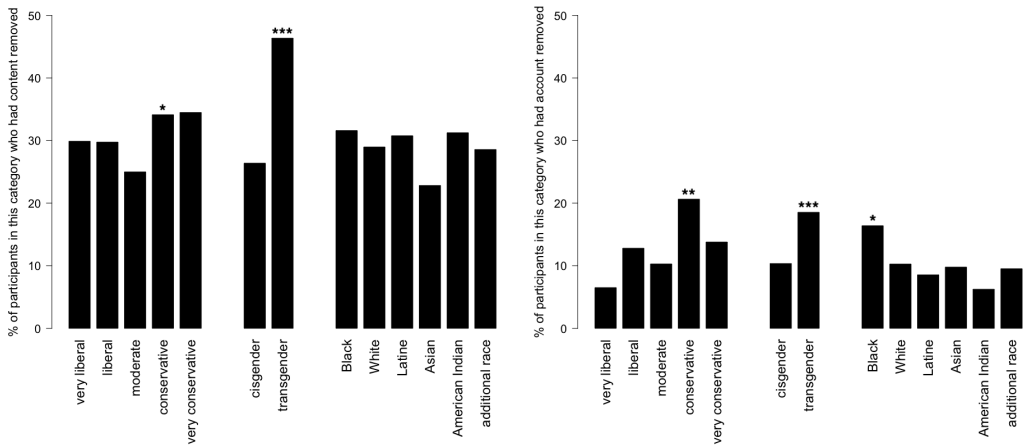

Fig. 1. Bar graphs showing percentages of participants who reported having social media content (left) and accounts (right) removed for reasons they disagreed with, categorized by political orientation, trans status, and race/ethnicity. Plots show that trans and conservative participants were more likely than others to have content removed, while trans, conservative, and Black participants were more likely than others to have accounts removed.

4.2 What Types of Content are Removed for Conservative, Trans, and Black Participants?

Table 5 displays results from our qualitative coding which determined types of content that participants described having removed on social media. The types of content commonly removed differed greatly between participant categories. Several types of content showed statistically significant effects or occurred entirely within only one category of participants (see Table 6). Table 7 summarizes our results.

4.2.1 Types of Content Removed More Frequently From Conservative Participants. Conservative participants were significantly more likely to have content removed that was deemed offensive or inappropriate by participants themselves or the platform's moderation mechanisms, related to the Covid-19 pandemic, including adult content or nudity that was against the platform's guidelines, clear instances of misinformation, and hate speech.

Conservative participants more frequently reported removal of content that we coded as offensive or inappropriate or allegedly so. For example, P236 described her removed Twitter post as "*It was disturbing image of a child,*" and later described it as "*bad images.*" Though she did not provide additional details, this seems to be describing actually disturbing content. P45 reported that Facebook removed a photo that he posted that "*resembles swastikas*," a visual that many would consider offensive or inappropriate. P243, who had content removed from Quora, wrote, "*I referred to some idiot as an 'idiot.' They told me that I violated their guidelines on civility. They deleted my post and made me click on a statement saying that I agree to abide by them.*" Some content deemed offensive or inappropriate related to criticizing or insulting others, such as P243's disagreement. Other participants suggested a potential bias from the platform and/or moderators leading to removal of content as inappropriate. P175 wrote, "*Twitter deletes my contents and wipes everything they term inappropriate.*" Unlike P243's response about a disagreement with another user, P175 did not describe a particular content removal instance, but instead noted a broader trend of content

Table 5. Counts of codes assigned to participants in each category (conservative, transgender, Black). Bold text designates differences between participants in that category and those not in that category. Cells in bold text with significance stars highlight codes significantly more present for participants in that category, according to the regression models in Table 5. Cells in bold text without significance stars highlight codes more present for participants in that category as indicated by complete or quasi-complete separation.

code	*n*	conservative	transgender	Black
n	207	70	66	100
types of content removed				
political content	45	17 (24.3%)	15 (22.7%)	21 (21.0%)
offensive/inappropriate or allegedly offensive/inappropriate	24	**12 (17.1%)***	5 (7.6%)	13 (13.0%)
content removed as adult despite following guidelines	23	2 (2.9%)	**14 (21.2%)*****	8 (8.0%)
Covid-related content	18	**13 (18.6%)***	4 (6.1%)	2 (2.0%)
adult content or nudity that does not follow guidelines	17	**10 (14.3%)***	3 (4.5%)	11 (11.0%)
misinformation (content clearly false)	17	**14 (20.0%)****	2 (3.0%)	3 (3.0%)
content insulting or criticizing dominant group (e.g., men, white people)	15	1 (1.4%)	**10 (15.2%)****	7 (7.0%)
content related to racial justice or describing racism	13	1 (1.4%)	4 (6.1%)	**11 (11.0%)***
hate speech (all types combined)	13	7 (10.0%)	1 (1.5%)	5 (13.0%)
copyright violation	12	0 (0.0%)	4 (6.1%)	8 (8.0%)
language / curse words	11	2 (2.9%)	3 (4.5%)	6 (6.0%)
content against online community's rules or norms	10	0 (0.0%)	3 (4.5%)	7 (7.0%)
self-promotional content (e.g., link, ad)	8	5 (7.1%)	0 (0.0%)	6 (6.0%)
hate speech: potentially racist	7	4 (5.7%)	1 (1.5%)	2 (2.0%)
queer content	6	0 (0.0%)	**6 (9.1%)**	0 (0.0%)
violent content	6	5 (7.1%)	1 (1.5%)	2 (2.0%)
content removed as misinformation but unclear whether it was	4	1 (1.4%)	3 (4.5%)	0 (0.0%)
hate speech: anti-LGBTQ+	4	2 (2.9%)	0 (0.0%)	2 (2.0%)
content removed as violent despite following guidelines	3	1 (1.4%)	**3 (4.5%)**	0 (0.0%)
feminist content	3	1 (1.4%)	0 (0.0%)	**3 (3.0%)**
hate speech: general or unspecified	3	**3 (4.3%)**	0 (0.0%)	0 (0.0%)
negative content	3	2 (2.9%)	1 (1.5%)	0 (0.0%)
self-referential slur	3	0 (0.0%)	1 (1.5%)	2 (2.0%)
trans content	3	0 (0.0%)	**3 (4.5%)**	0 (0.0%)
trans surgery content	3	0 (0.0%)	**3 (4.5%)**	0 (0.0%)
account-related issues				
spam account or alleged spam account	8	2 (2.9%)	1 (1.5%)	6 (6.0%)
security issue	7	4 (5.7%)	1 (1.5%)	6 (6.0%)
account hacked	6	3 (4.3%)	1 (1.5%)	4 (4.0%)
impersonation or alleged impersonation	4	0 (0.0%)	1 (1.5%)	3 (3.0%)
"real name" issue	3	1 (1.4%)	2 (3.0%)	1 (1.0%)
interpersonal issues				
interpersonal conflict	8	4 (5.7%)	1 (1.5%)	5 (5.0%)
disagreement in online community	7	1 (1.4%)	**4 (6.1%)***	2 (2.0%)
content mass-reported (by multiple people/bots in targeted campaign)	3	1 (1.4%)	2 (3.0%)	1 (1.0%)
perceptions				
perception that others' similar content allowed while theirs removed	20	5 (7.1%)	7 (10.6%)	8 (8.0%)
perceived anti-conservative bias	10	**10 (14.3%)**	0 (0.0%)	0 (0.0%)
perceived censorship	10	7 (10.0%)	2 (3.0%)	1 (1.0%)
perceived anti-trans bias	6	0 (0.0%)	**6 (9.1%)**	0 (0.0%)
perceived anti-people of color bias	5	0 (0.0%)	3 (4.5%)	3 (3.0%)
perceived anti-queer bias	5	0 (0.0%)	**5 (7.6%)**	0 (0.0%)
perceived anti-Left bias	3	0 (0.0%)	2 (3.0%)	1 (1.0%)

Notes: Table only includes codes that were assigned to three or more participants. The "conservative" category includes all participants who stated they were either conservative or very conservative. People could be in multiple categories, so row numbers may add up to greater than the total *n* for each code.

Table 6. Logistic regression models examining social media content moderation codes associated with participants being conservative (Model 3), transgender (Model 4), and Black (Model 5).

	Conservative (Model 3)			**Transgender (Model 4)**			**Black (Model 5)**		
Variable	Coef. (SE)	OR	95% CI	Coef. (SE)	OR	95% CI	Coef. (SE)	OR	95% CI
intercept	-1.19*** (0.24)	0.30	(-1.68,-0.74)	-1.11*** (0.19)	0.33	(-1.49,-0.75)	0.01 (0.16)	1.01	(-0.31,0.32)
offensive/inappropriate content	**0.96* (0.48)**	**2.61**	**(0.00,1.91)**	-	-	-	-	-	-
adult content against guidelines	**1.32* (0.59)**	**4.73**	**(0.65,2.50)**	-	-	-	-	-	-
removed as adult but follows guidelines	-1.16 (0.78)	0.31	(-3.04,0.17)	**1.55*** (0.47)**	**5.08**	**(0.73,2.57)**	-	-	-
violent content	2.02† (1.16)	7.56	(0.00,5.04)	-	-	-	-	-	-
Covid-related content	**1.31* (0.64)**	**3.70**	**(0.06,2.61)**	-	-	-	**-1.59* (0.72)**	**0.20**	**(-3.51,-0.19)**
misinformation (content clearly false)	**2.14** (0.71)**	**8.47**	**(0.84,3.70)**	-	-	-	-0.71 (0.72)	0.49	(-2.29,0.64)
self-promotional content	1.60† (0.82)	4.97	(0.03,3.36)	-	-	-	-	-	-
hate speech	-	-	-	-2.19† (1.15)	0.11	(-5.23,-0.33)	-	-	-
related to racial justice or racism	-1.41 (1.14)	0.25	(-4.42,0.49)	-	-	-	**1.63* (0.78)**	**5.08**	**(0.26,3.52)**
content criticizing dominant group	-1.56 (1.15)	0.21	(-4.58,0.31)	**1.96** (0.60)**	**7.07**	**(0.83,3.23)**	-	-	-
perceived censorship	-	-	-	-	-	-	-1.64 (1.09)	0.19	(-4.59,0.17)
security issue with account	1.50† (0.85)	4.46	(-0.17,3.30)	-	-	-	1.65 (1.10)	5.19	(-0.18,4.61)
disagreement in online community	-	-	-	**2.02* (0.92)**	**7.54**	**(0.30,4.10)**	-	-	-
AIC	231.31			240.26			271.71		
n	207			207			207		

† p <0.10; * p < .05; ** p < .01; *** p < .001

Note: The "conservative" category includes all participants who stated they were either conservative or very conservative.

deletions for posts Twitter's policies deemed as inappropriate. In some of these cases we do not have enough information to determine whether the content violated site policies or not, and some may represent content moderation gray areas. Others (e.g., the "disturbing image of a child") seem to be clear cut removal decisions.

Conservatives reported content removals for misinformation more frequently than others. Some of this misinformation related to U.S. politics and marginalized communities. P217's content was removed from Facebook and TikTok for "*crude content and false information.*" The removed content was "*content involving the presidential candidates and LGBYT [sic] community.*" In our dataset, most conservative content removal experiences due to misinformation were related to the Covid-19 pandemic; content involved criticism or support of government responses to Covid-19 such as attempts to open schools in the midst of the pandemic. For example, P239 had shared, "*a post that stated that when children return to school that 2 percent of them will catch the virus.*" Many conservative participants who shared Covid-19 information like P239 reported that it was removed for "*misinformation,*" "*false facts,*" or other similar descriptors. Others criticized medical responses such as treatment of Covid-19. P265 posted a "*video by numerous doctors discussing the pandemic. They disagreed with the press negative portrayal of using Hydroxychloroquine*" on Facebook. P105 associated Covid-19 information with other political content and said, "*I have recently shared posts about Covid and political views on Facebook and they were all taken down for 'false facts.*"' These quotes demonstrate misinformation's prevalence on social media, especially related to health-related content and specifically posted by conservatives. Moderate and liberal participants described far fewer instances of content being removed as misinformation. Yet these data also highlight differences between what people with different political orientations believe is true and false: because our survey specifically asked about content removal decisions that participants disagreed with, it follows that many participants who posted untrue content believed that that content was true and should be allowed to circulate online. Other participants may not have believed that their content was true (e.g., they may have intentionally posted misinformation to "troll" people or communities), yet may have nonetheless believed such content should stay online. We cannot fully speak to participants' motivations for posting misinformation, but our findings suggest

that conservatives were more likely than others to believe the misinformation should remain on social media platforms, regardless of the sincerity of their intentions. Based on our understanding of Covid-19 facts and science, we were able to clearly categorize many of these data points as misinformation rather than content moderation gray areas.

Adult content or nudity not following platform guidelines also appeared more in conservatives' survey responses. P326 described YouTube's removal of "*a video about pornography and link I sent to my status for my friends to follow and link and view.*" In addition to pornographic content, one conservative participant experienced removal of a photo posted on Instagram: P319 wrote, "*I posted a picture with a little boob action and it was taken down sadly.*" These quotes describe instances where participants posted adult content that was clearly not aligned with site policies.

Hate speech was the final type of content removed more frequently for conservative participants. When asked about the type of content removed, P244 said, "*I said I hate LGBT and they removed it and banned me for 30 days.*" According to P244, when Facebook removed this content, "*they said it was hate speech it wasn't it was just truth.*" While Facebook correctly categorized anti-LGBTQ+ content as hate speech, P244 disagreed and described that he should be able to post about his hatred of LGBTQ+ people. This example demonstrates that people of different political orientations do not agree with platform decisions about what constitutes hate speech. P178 described a Reddit community where people would say "*harmful or hateful things.*" Eventually his account was removed for hate speech because, "*I think they just did not want the community, that these hateful words were taking place in, to remain on Reddit.*" While P178 did not consider his own behavior to be hate speech, his membership in a community where hate speech was prevalent eventually led to his content and account being removed. Most of conservatives' data in this category clearly violated site policies on hate speech.

The types of content removed more frequently from conservative participants – (allegedly) offensive content, misinformation, adult content, and hate speech – primarily represent harmful content that sites removed to cultivate safe online spaces with accurate information, rather than falling into content moderation gray areas. People may argue that some of this should be allowed on social media, and clearly the participants in our study disagreed with these removal decisions. However, at least according to site guidelines and standards of common decency, many of these removal decisions appear to be relatively clear cut.

4.2.2 Types of Content Removed More Frequently From Trans Participants. Trans participants were more likely to have content removed that was classified as adult despite following the site's guidelines (including trans surgery content), content that insulted or criticized a dominant group (e.g., men, white people, cisgender people), content that they considered queer or trans, and content that was removed as violent despite following guidelines.

Trans participants reported more experiences where content was removed as adult despite following platform guidelines according to participants' descriptions of the removed content and our analysis of these descriptions as compared to site guidelines. P106 described their experience following platform guidelines but still facing removal of content: "*I posted a selfie on Tumblr where I was not wearing a shirt, but had my chest covered by tape so that it appeared flat and no tissue was visible. Tumblr took it down despite it being allowed.*" P106 took issue with how the content removal sexualized and classified their body, stating that "*trans bodies aren't inherently pornographic and do not need to be policed that way.*" With this group of participants, content removed as adult often included content related to gender affirmation surgeries. For example, P189 described, "*I shared a link to a post op phalloplasty blog that has posts that have nude images. The blog is really for educational purposes but I assume Facebook took it down because the link had an thumbnail preview. I don't believe the thumbnail itself had nudity.*" This experience made P189 "*annoyed*," because "*I*

was trying to share a link to help educate others." It should be noted that Facebook explicitly allows nudity in the context of "health-related situations, for example, gender confirmation surgery" [28]. While surgery photos are not actually a content moderation gray area according to policy because they are explicitly allowed, they become a gray area in content moderation enforcement because, without the necessary contextual information (e.g., that the photo was posted in a trans support group), they may appear similar to other nude photos to both human and machine moderators. Not all experiences were related to surgeries and images. For instance, for P86, "*a post about being gay was removed as 'adult content'*" on Tumblr. Some platforms mistakenly deemed LGBTQ+ content "adult," leading to content removals. Some trans participants experienced "adult" content removals related to art, such as P14's "*post containing images of art with artistic non-sexual nudity*" which "*Tumblr's bots probably identified... as adult despite following the site's rules which are supposed to allow artistic nudity in certain contexts.*" Finally, P101 had "*multiple cat pictures of all things flagged and removed... Apparently videos of my literal cats playing with feathers and napping and such is 'inappropriate sexual content.'*" Trans participants' experiences with misclassified "adult" content removals ranged from surgery content to cat photos, but were similar in that they were removed despite not actually violating site policies, which decreases people's agency to express their identity online.

Criticisms of a dominant group such as men or white people were prevalent in trans participants' content removal experiences. P118 described their content removal on Facebook: "*I posted 'Men are trash' in reference to my sexual assault,*" an experience echoed by P192. Men were not the only dominant group criticized. P192, a trans person of color, also experienced a content removal when a post was considered critical of white people: "*I had reposted some meme or picture of a tweet, something along the lines of 'don't invite me to all-white LGBTQ events, those are just gay KKK rallies.'*" This participant attributed their content removal to this criticism of racism within the LGBTQ+ community and their specific criticism of white people within this community. To P192 this was frustrating, because "*Facebook is very hypocritical since it will find nothing wrong if you report content that contains, say, anti-Black comments...[or] neo-Nazi comments... Yet Facebook will be quick to remove stuff that says something negative about white people.*" Several participants also discussed removal of content criticizing transphobic people and Trans-Exclusionary Radical Feminists (TERFs)[4]. P21 responded, "*I received a brief suspension from Twitter and had a post removed by them for telling a TERF (Trans Exclusionary Radical Feminist, i.e. a transphobe) to 'fuck off.' Legitimately, the entire content of the post was telling this transphobe to 'fuck off,' which evidently set off some sort of Twitter protocol that led to the removal of the post and the brief suspension.*" Other trans participants shared similar content removal experiences after criticizing TERFs. While posts like these might offend people in dominant groups such as men, white people, and cisgender people, criticism of identity-based categories is different when coming from a dominant group (e.g., anti-trans rhetoric from cisgender people) than when coming from a marginalized group (e.g., anti-TERF content from trans people) given the power imbalance – thus representing a content moderation gray area. However, social media sites often treat both types of content similarly when it comes to content moderation, which can make online spaces hostile for trans people.

Trans participants also more frequently experienced removal of trans and queer content. P98 experienced a content removal related to "coming out" as trans on Facebook: "*I had come out as transgender, and the post was taken down within an hour.*" This was distressing to P98, as "*it made me remember yet again that trans people don't get a place on social media.*" P86 described how Tumblr "*falsely applied their algorithms to remove content about lesbians,*" which they considered

[4]TERFs, which stands for trans-exclusionary radical feminists, often harass and exclude trans people, particularly trans women, both online and in physical spaces [58, 124]

"*homophobic.*" Similarly, P173 Facebook's account was taken down when she "*made an innocent post about bisexuals.*" On TikTok, P93 experienced "*videos and posts removed with little explanation as to the reasons. The claim was that they violated regulations, but they only included non explicit and inoffensive LGBT content.*" Each of these examples describes content that was seemingly wrongfully removed (rather than representing a conent moderation gray area), yet that related to participants expressing their personal identities as trans and/or LGBTQ+ people.

In this section, we described types of content that were more frequently removed from trans participants, including content that was removed as adult despite following platform guidelines, content that criticized dominant groups, and trans/queer content. Each of these represents content that was related to participants expressing their marginalized identities as trans people, and much of it was mistakenly removed despite the fact that it technically aligned with site guidelines, or represented content moderation gray areas.

4.2.3 Types of Content Removed More Frequently From Black Participants. Black participants were more likely to have content removed if it related to racial justice or described racism, or content describing feminist viewpoints.

Black participants frequently reported removal of racial justice content or content describing racism. P29 described their experience on Facebook: "*I was discussing anti-racism and my post got removed for hateful speech because because I mentioned 'white people' in a 'negative' way.*," which was echoed by several other Black participants and relates to the "criticizing dominant group" code we described in the previous section. P29 continued: "*Facebook doesn't like when people use 'white people,' 'whites,' or 'yt people' in a negative context, regardless of if that negative context is about discussing the role of 'whiteness' in racism/white supremacy.*" Instances like these seem to represent content moderation gray areas, where human and machine moderators have difficulty distinguishing between content that is allowed and content that is not, since what is and is not hate speech may depend on the poster's identity and the identity of those described in the post. P147 described another experience where content related to racial justice and racism was removed: "*I made a post about actual, factual atrocities that Black people have suffered from white supremacists and systematic racism in America. I said it's time for white people to do their part if this country is ever going to heal. That they need finally and forcefully confront their racist family and friends because Black people can't fix a problem that we didn't create.*" For P147, the content removal "*reminded me that Facebook is not a safe space for women and Black people. I found a safer platform.*" When Tumblr removed her content about Black Lives Matter, P154 stated that the experience "was incredibly annoying and hurtful. I put thought into writing the post and having it be taken down like that stung." These quotes demonstrate how content removals related to racial justice content substantially negatively impact the Black social media users who post that content, and decrease their agency and ability to express their experiences with racism online. It is unclear why these types of posts were removed, given that they do not seem to violate site policies.

Black participants also were more likely than other groups to report removal of content considered feminist. P130 wrote, "*I uploaded a post about gender equality and sexual harassment on a group I belonged to, the post was taken down.*" Most of the removed feminist content related to sexual harassment. P4 responded, "*My Facebook account was taken down because they said 'I was being rude to a fellow user and they don't support rudeness.' All I did was speak against men slut shaming women all the time or coming to their women to send dick pictures and harass women.*" Actions deemed "rude" or unacceptable by platforms were viewed as feminist or socially just by participants. Data in this category seem to be either wrongfully removed despite following site policies, or may represent content moderation gray areas as platforms decide which types of content are and are not appropriate when critiquing sexism and harassment of women.

Table 7. Results summary.

Group	Types of content removed more frequently than for other groups	Broader trend
conservative participants	(allegedly) offensive/inappropriate, misinformation, Covid-related, adult, hate speech	harmful content often removed to cultivate safe spaces and accurate online information, usually against site policies
transgender participants	removed as adult despite following site guidelines, insulting/criticizing dominant group, trans content, queer content	content related to expressing participants' marginalized identities, usually follows site policies or falls into gray area
Black participants	racial justice or describing racism	content related to expressing participants' marginalized identities, usually follows site policies or falls into gray area

We described several types of content that were more likely to be removed from Black participants in our study: content related to racial justice or racism, and feminist content. These tend to be very different from the types of content removed from conservative participants. Content about racial justice was often especially meaningful for participants, as it related directly to their identities as Black people and was attempts to make important points about systemic racism to their social media audiences. Removing racial justice content as "hate speech" because it critiques white supremacy falls into a gray area with respect to platform policies (which have recently been adjusted on Facebook [26]), yet such policies should be reevaluated broadly as they negatively impacted Black participants and limited their online participation.

4.3 What Types of Accounts-Related Issues Occur for Conservative, Trans, and Black Participants?

In Table 5 we also show qualitative coding results related to account takedowns. Participants described account removals that occurred related to accounts being spam or alleged spam, security issues, account hacking incidents, impersonation or alleged impersonation, and "real name" issues. Unlike the differences we saw in types of content removed, account-related issues did not show significant differences between groups. P128, a trans participant, described barriers to using Facebook and Oculus related to "real name" policies: "*Because Facebook purchased Oculus, I had to tie my Facebook account to my Oculus account in order to access some features. When I attempted to make a Facebook account with my chosen name, it was immediately deleted because I couldn't produce a valid ID with the name on it. Additionally, Facebook has no way for an individual to change their real name on the Oculus account, which is a problem for many trans individuals I know.*" While "real name" issues did happen to several trans participants like P128, these issues also occurred for at least one Black and one conservative participant. P146, a conservative participant, responded, "*The memorable take down of my account according to the platform, they said that they suspect that I was using a fake name or information.*" Unlike P128, P146 did not associate this account removal experience with a component of their identity. While we do not see significant differences between groups related to account-related issues, when participants' accounts were removed, it was often related to the types of content they had posted. Thus, the differences we found in account removals (in Table 4 and Figure 1) stem from content-related differences between groups as described in section 4.2.

4.4 What Perceptions do Conservative, Trans, and Black Participants Hold About Their Removed Content and Accounts?

In the last section of Table 5, we present results of our qualitative coding related to perceptions participants held about their content or account removal experiences. Here, we see several differences between groups: conservative participants were more likely to perceive anti-conservative bias from social media platforms or moderators, while trans participants were more likely to perceive anti-trans or anti-queer bias from social media platforms or moderators.

4.4.1 Perceptions Held By Conservative Participants. Conservative participants more frequently perceived that platforms and their content moderation policies involved anti-conservative bias. P224 argued, "*Almost anything conservative I post now is fact checked or removed.*" Similarly, P223 wrote, "*They censor conservative things but not liberal.*" After posting anti-LGBTQ+ content, P244 said, "*They [Facebook] said it was hate speech it wasn't it was just truth. Because I am a conservative so they take down anything that goes against their agendas.*" These quotes about anti-conservative bias on social media echo what we see in right-wing media sources like Fox News and Breitbart [71], or rhetoric from the past presidential administration, and may have in fact been influenced by these.

Many conservative participants justified their perceptions of anti-conservative bias using removal of Covid-19 content and/or Donald Trump-related content as examples. P107 responded, "*My Twitter was temporarily taken down. Again my pro-Trump videos showing the truth behind of Covid were being taken down, so I waited and kept putting them back up, finally Twitter just suspended my account and won't let me use it if I keep on putting my Covid videos.*" P219 also described her perception of a social media platform's anti-conservative bias related to Trump: "*Facebook hates Trump. Biased.*" P107 attributed Facebook's alleged anti-conservative bias to "*just that the left is controlling most social media platforms.*" For some conservative participants, perceptions of social media platforms as biased against conservatives stemmed from a perception of those platforms being controlled by liberals.

4.4.2 Perceptions Held By Trans Participants. Trans participants more frequently perceived that social media platforms held both an anti-trans bias and an anti-queer bias. P98 described his perception of Facebook: "*Facebook has a history of deleting trans people's profiles and making them need to give their dead names in order to continue using their platform.*" According to P176, Facebook "*intended to censor lesbian advocacy.*" P128 experienced issues with their account name and described his perception of the "real name" incident he experienced with Facebook and Oculus and its impact on trans people being able to use both systems: "*While Facebook is likely attempting to reduce spam accounts, they are disallowing trans individuals from using many of their products. By not making exceptions for trans individuals and making the name on an Oculus account permanent,... they may very well be making a political message.*" Content and account removal experiences like these led participants to perceive that platforms and moderators held anti-trans and anti-queer bias. Several trans participants identified anti-trans bias that upholds the transphobic rhetoric of TERFs while preventing trans users from defending themselves online. P64 described this difference as, "*Twitter unfairly moderates trans users vs TERF users.*" According to participants, platforms quickly removed their content criticizing TERFs or defending themselves and other trans people against transphobic content, while allowing TERF content and accounts to remain online. P21 wrote, "*I think, more than anything, I was annoyed that Twitter takes more actions like this against people like me who are just trying to defend ourselves than people like the TERF in question who inflict active harm on a daily basis and aren't called out by Twitter at all for it.*" Platforms' perceived support of TERFs likely influenced some participants' perceptions of anti-trans bias related to content moderation. These

quotes support previous literature that has shown ways that social media sites can be particularly difficult places for trans users [52, 104].

4.4.3 Perceptions Held By Black and Trans Participants. Black participants and trans participants (including two trans people of color and one white trans person) were more likely than conservatives to perceive an anti-people of color (POC) bias in their content moderation experiences. As we presented in section 4.2.3 in relation to criticisms of a dominant group, P29 perceived Facebook as a platform that censors negative portrayals of white people, and attributed these content removal decisions to the platform's upholding of white supremacy and racism. P147 and P68 explicitly identified how they perceived anti-POC bias in content moderation to impact Black users. P147 wrote, "*Facebook has a tendency to protect white privilege while silencing Black voices and women. I've had content removed addressing misogyny too. It reminded me that Facebook is not a safe space for women and Black people. I found a safer platform.*" P68 described an instance when Facebook removed a post about George Floyd's murder as "hate speech." When asked why they thought the content was removed, P68 shared, "*the same reason I made the post in the first place, anti-Blackness.*". These perceptions from Black and trans participants indicate how social media platforms' removal of content related to racial justice and racism resulted in users viewing those platforms as biased and unwelcoming for people of color.

4.4.4 Perceptions Not Significantly Different Between Groups. As noted in the previous sections, each group described perceptions that platforms were biased against their particular identity (e.g., conservative, trans, queer, POC). In this sense, we note a similarity between groups in that they all felt oppressed by platforms that they considered to be biased against them. As another similarity, there was no significant difference between groups regarding the perception that others' similar content was allowed while theirs was removed; all groups equally described this perception. P239, a "very conservative" white woman, had content about Covid-19 removed from Facebook. She wrote, "*It did offend me. I have seen much worse posted and stay up.*" Participants with other identities shared similar sentiments about "worse" or "more harmful" content staying up when their own content faced removal. P33, a "very liberal" white man, wrote, "*It made me annoyed and more hostile towards Instagram for removing my post and allowing much more offensive and directly harmful posts to stay up.*" P161, a politically moderate Black woman, said, "*Facebook allows some of the most disgusting things to be posted but removes innocent posts.*" Across different political orientations, races, and genders, participants perceived content removals as unfair because they viewed content remaining online they deemed as more inappropriate or harmful than their own removed posts. Additionally, there was no significant difference between perceived censorship. P278, a "very conservative" white woman, had content related to Covid-19 removed from Facebook and YouTube, and wrote that the sites were "*censoring my opinions.*" Similarly, after being blocked from Twitter, P174, a liberal white woman, stated that Twitter was "*censoring my contents.*" Despite vastly different political orientations, participants across the political spectrum similarly perceived that they were being censored by social media platforms.

5 DISCUSSION

We have examined three groups who reported experiencing content moderation more often than others – conservatives, trans people, and Black people – and described differences in the types of content each had removed from social media sites. Conservative participants reported content removals for types of material that platforms have an interest in removing to promote safety and accuracy: offensive/inappropriate content (or allegedly so), misinformation, Covid-related content, hate speech, and adult content. Much of this content appeared to be in violation of platforms' policies. On the other hand, participants with marginalized identities – trans and Black participants

– reported content removals related to their personal identities (e.g., specifically trans and/or queer content, content related to racial justice or racism), or content that appeared to be removed despite following site guidelines or that fell into content moderation gray areas. Thus, while these three groups may have faced similar levels of content and account removals, types of content removals for each group were substantially different, and demonstrated different stakes in terms of personal agency to express one's identity online.

Yet as any content moderation research must acknowledge, what is considered worthy of removal (i.e., what is and is not designated as offensive, inappropriate, misinformation, and hate speech) differs greatly depending on who you ask. As Jillian York wisely observed, "although censorship as a concept is value-neutral, it is all too often used only to describe the restrictions of which we disapprove" [125]. By virtue of the way we asked our survey question about content and accounts removed "for reasons you disagree with," every participant in our study disagreed with the content or account moderation decision a platform or moderator made. Every participant believed that their content or account should have remained on the site. Thus, when we (as researchers) say that hate speech should be removed from social media sites, we know that people in our dataset disagree, and many even disagree on what is and is not hate speech (as demonstrated by P244 who stated that anti-LGBTQ+ content was truth, not hate speech). Similarly, when we say that misinformation should be taken down, and when sites like Facebook and Twitter agree [34, 118], these positions run counter to the beliefs of the participants in our dataset who posted clearly false content and disagreed with its removal. While we cannot resolve these disagreements in this paper, our work in providing empirical evidence for whose social media content is disproportionately removed, and potential insights into why, provides a first step that we hope future research and policy can build on.

While conservatives frequently claim to be victims of disproportionate censorship on social media sites (e.g., on conservative venues like Fox News and Breitbart [71]), previous studies found no evidence to support these claims [80, 88]. In this way, our results differ from previous research because we show that conservatives are actually significantly more likely to experience content removals than people with other political orientations. However, by examining what types of content conservatives frequently reported having removed from social media sites, we found that a reason for their disproportionate content removals was that they frequently posted content violating site policies. This is in stark contrast to the other two groups we examined – trans and Black people – whose social media content removals were more likely to be content that either was removed despite following site policies, or that fell into gray areas with respect to policy and enforcement. Based on our results, we can highlight conservative content removals as instances in which content moderation *seems to be actually working as it should*. Rather than content moderation false positives or gray areas, the conservative content removals in our dataset were more likely to represent true positives: content that violated site policies, and thus was correctly removed.

However, whether something was "correctly" or "incorrectly" removed depends on a site's policies, and our results show that conservative participants frequently perceived platforms and their content moderation policies to perpetuate anti-conservative bias. While our study does not specifically examine removals of entire online communities, it is worth considering such cases, as they represent massive content removals that can reinforce perceptions of platform bias, which can impact community members' online activities and pose consequences on information ecosystems. Studies have found that while banning toxic communities from a mainstream platform may reduce elements like hate speech on that platform [12], conservative communities that were banned from mainstream platforms, such as r/The_Donald, may migrate to other platforms with more lax policies or set up their own dedicated sites [97]. In this way, exiled communities move hate speech and misinformation to environments more amenable to it. Indeed, we recently saw a massive

conservative migration to Parler, an app (now suspended for the risk of further inciting violence [39]) describing itself as the world's "premier free speech social network," after Facebook and Twitter began more aggressively removing misinformation [64]. Additionally, the new Trump-supported social media platform Gettr [89], targeted at conservatives, involves moderation policies much more lax than those we see on mainstream sites like Twitter, Facebook, and Reddit. As such, we may continue to see online political polarization that depends on how platforms moderate the types of content that we found conservatives often have removed: misinformation, hate speech, and content considered inappropriate on mainstream platforms.

Our results provide empirical evidence for some of the ways social media content moderation silences marginalized groups like Black and trans people. Gerrard and Thornham [41] described social media platforms' prescriptive power in a feminist context by introducing the concept of 'sexist assemblages,' which describes how social media content moderation "perpetuate[s] normative gender roles, particularly white femininities, and police[s] content related to women and their bodies." Assemblages, drawing from Deleuze and Guattari [21] and Bucher [10], refers to social media platforms' dynamic and complex content moderation processes which bring together multiple elements, including both algorithmic and human techniques, to impose policy. Gerrard and Thornham [41] posited that similar silencing likely also occurs for other marginalized social media users, and our work provides evidence for this claim and describes how trans and Black people experience such anti-trans and racist assemblages. While conservatives also experience silencing and perceive anti-conservative bias on social media sites, there is a vast difference between silencing conservatives' misinformation and hate speech and silencing trans and Black users' personal identity-related content.

5.1 Potential Ways Forward for Equitable Content Moderation for Marginalized Social Media Users: Embracing Content Moderation Gray Areas

Because content moderation and platform moderation processes are largely invisible [43], understanding how to make changes that will benefit marginalized groups is challenging. A first step is documenting marginalized people's experiences with content moderation, which we do in this work, adding to evidence provided by the Electronic Frontier Foundation [35] and community-based groups like Hacking/Hustling [6, 7] and Salty [102].

Some have proposed jury-based [32, 128] or advisory-board driven governance approaches like Facebook's Oversight Board [8, 72] as solutions to thorny content moderation issues; however, such approaches will likely fall short when it comes to gender and racial minorities' disproportionate content removals. In our study, trans participants had content removed at a higher percentage than any other group; yet, the trans population in the U.S. is less than 1% [70], so the chances of a jury or advisory board including a trans member are low. The chances of that governing body including a trans person of color, or multiple trans people to represent the diversity of the trans community, is even lower. Thus, even if social media platforms were to invoke a "jury of one's peers" when making content moderation decisions, that jury may likely still decide to remove content criticizing a dominant group, may mistake in-group self-referential slurs for hate speech, and may wrongfully remove trans content as adult even when it follows site guidelines.

Platforms may also support marginalized groups by increasing transparency and accountability, and by providing more explanation and detail to help users understand content moderation policies [17, 42, 85]. While helpful for all users, these approaches take more of a "one size fits all" approach, which will privilege some users while continuing to marginalize gender and racial minorities. Thus, these approaches do not fully address marginalized populations' experiences with disproportionate content and account removals.

Making social media content moderation more equitable for marginalized users may involve considering more radical approaches. For example, Mulligan et al. [84] introduced the concept of "contestability" – the ability for users to meaningfully challenge algorithmic determinations – which could help in cases when marginalized people's content is removed. Such an approach would need to go far beyond current social media appeals processes, which are often insufficient [122]. Further, Christian et al. [15] argued that platforms should be reconceptualized entirely to center communities rather than rely on corporate power structures, and Haimson et al. [52, 53] have advocated for cooperatively-driven, community-based platforms built with and for particular marginalized communities. Related to nudity and adult content, Spišák et al. [112] suggested that social media platforms could focus on gaining viewers' consent before they access such content, rather than censoring it outright. Such shifts could enable more equitable futures, but will take substantial work and time.

One step that could help to mitigate harm more immediately, within corporate power structures, is for platforms to design to embrace content moderation gray areas (i.e., content where it is unclear whether it should or should not be removed). Roberts [99] described how, despite the complexity and difficulty of content moderation decision-making processes, there are only two possible outcomes: remove or keep up. What if there were more options?

5.1.1 Approaches for Embracing Content Moderation Gray Areas. Content moderation gray areas are by definition complex and tricky to moderate, causing difficulty for both human moderators and computational moderation systems. We provide several ideas for considering how to embrace content moderation gray areas' complexity rather than trying to force content into categories they resist.

1) Apply tags or blur content that may be inaccurate or explicit. One way to moderate gray areas has already been implemented widely on sites like Facebook, Twitter, and Reddit. Rather than removing potentially false content entirely, these sites often clearly label misinformation or contested content so that users can take caution before clicking. This is especially important because misinformation can be dangerous and harmful, especially in health-related contexts like the Covid-19 pandemic, and particularly for vulnerable populations who may lack information literacy. Similarly, sites often blur violent or adult content so that a user must take the extra step of choosing to view that content and clicking on it, rather than such content appearing in one's timeline automatically [52]. These types of approaches are widely in use already, and represent a promising step forward for embracing content moderation gray areas and providing a mechanism beyond binary allow / remove decisions. Tags and blurred content are useful both for the types of content we found that conservatives participants posted (e.g., misinformation) and the types of content we found that trans participants posted (e.g., content removed as adult despite following guidelines).

2) Apply different moderation approaches for different online spaces depending on context (e.g., timeline vs. private group). Moderation approaches often treat all content across a site similarly, despite vast differences in the audiences for whom that content is posted. As one example, trans surgery content is allowed on most social media sites, but that does not mean that posters would want such content to be viewed by everyone in their network. Rather, such content is often shared in private online communities of similar others. Yet moderation approaches often treat this content similarly to content that is meant to be broadcast widely. While some sites (e.g., Reddit, Discord) already employ different moderation approaches for content in specific online communities [13], all sites should employ different moderation approaches for different online spaces depending on context and audience, especially considering online community contexts [122]. As part of this approach, sites must allow marginalized users to establish particular online spaces where certain types of gray content is allowed to stay up. Applying different moderation approaches for different

online spaces where marginalized individuals congregate aligns with calls to increase how much context is considered during content moderation decisions [11]. We do not advocate for applying more lax moderation approaches in private online spaces for the types of content conservative participants frequently had removed, as this would allow misinformation and potentially harmful content to spread in private spaces, which could increase online polarization and radicalization (as we have seen with sites like Parler and Gettr [18, 89]).

3) Develop specialized tools especially for particular marginalized groups. Jhaver et al. [67] suggested that sites should develop specialized tools that can better meet the needs of marginalized groups like trans people when experiencing online harassment and abuse. Similarly, knowing that trans people and Black people are more likely to have content removed from social media sites, specialized tools to moderate trans and Black users' content and accounts could help to decrease these disparities. Future research should conduct design research with trans and Black social media users to determine how such tools and systems should be designed.

4) Involve marginalized communities in creating moderation policy. Doing the work to embrace gray areas will require community involvement. To learn how to reduce content moderation disparities between trans and cisgender social media users, and between Black and white social media users, social media policy managers should involve trans and Black people in forming policy. This could involve hiring members of these groups, bringing them on board as consultants, and/or conducting in-depth research to learn more about their experiences and how policy could evolve to be more equitable.

We have outlined several approaches for embracing content moderation gray areas rather than attempting to improve accuracy by removing or reducing gray areas. Further, we align with Vaccaro et al.'s [122] call to move toward human-centered approaches to designing content moderation policies and appeal processes. Rather than attempting to push content moderation gray areas to one side of the line or the other and quickly make a remove / keep up decision, platforms can embrace gray area content and the users who create it by leaning into and directly addressing the complexities.

5.2 Limitations

Like any survey-based research, this work involves limitations. First, due to response bias [96], those who responded to our survey are not representative of all those who have experienced social media content and account removals. Next, as noted in Methods, in the absence of being able to observe content and account removals directly, we relied on participants' self-reports of moderation incidents, which may not be entirely accurate. Next, this research was conducted in a U.S. context, and may not generalize to non-U.S. and non-Western contexts; future research should examine marginalized groups' content and account removal experiences in other geographic and cultural contexts. Additionally, our data collection time period (Fall 2020) greatly influenced participants' responses and their removed content. In particular, the Covid-19 pandemic, 2020 U.S. elections, and reactions to the murders of Breonna Taylor, George Floyd and other Black Americans appeared frequently in participants' survey responses. Other experiences might not be as represented, especially for Black and conservative participants; however, these societal events provide an interesting backdrop to understand content moderation in the 2020 U.S. political climate. As another potential limitation, participants demonstrated some confusion regarding platforms' content moderation processes, and sometimes made incorrect assumptions about how or why their content faced removal. For example, a participant might think that content removed from a specific subreddit on Reddit was removed by the platform, rather than by subreddit moderators unaffiliated with the platform itself. However, misconceptions like these provide unique insight

into participants' content moderation experiences. In future work, we will use interviews to further interrogate participants' assumptions about content removal processes. Finally, we consider it a limitation that our work did not center sex workers, who are a primary target for disproportionate social media removals and face some of the most dire consequences [6, 7]; while our dataset may have included sex workers, we did not ask participants about this directly.

6 CONCLUSION

We have contributed a mixed methods empirical study highlighting three groups who experience disproportionate levels of social media content and account removals: political conservatives, trans people, and Black people. These findings echo claims commonly heard in the news about particular groups facing censorship on social media. However, we show how the *types* of content removed from each group, and the removals' *implications*, are substantially different for marginalized groups such as trans and Black participants. For trans and Black social media users, content removals limit their ability to post content related to their marginalized identities, and thus to participate in the public sphere. For conservative participants, on the other hand, content removals often demonstrate enforcement of site policies intended to remove harmful and inaccurate content. Knowing that marginalized social media users frequently experience content and account removals that limit their online participation, we advocate for social media sites to embrace content moderation gray areas, such as by involving marginalized people in forming policy and by designing moderation tools specifically for marginalized groups. We look forward to future social media content moderation research and design that explores how to take these next steps.

ACKNOWLEDGMENTS

We thank the participants in this study for sharing their experiences with us. Thanks to colleagues for feedback on our paper and survey, and for helpful conversations which helped inform this paper: Ben Zefeng Zhang, Claire Fitzsimmons and the Salty Algorithmic Bias Collective, Hibby Thach, Jane Im, Josh Guberman, Julie Hui, Linda Huber, Nazanin Andalibi, Patrick Carrington, Rahaf Alharbi, Robin Brewer, Samuel Mayworm, Shakira Smith, Shanley Corvite, Sarita Schoenebeck, and Silvia Lindtner. Thanks to anonymous reviewers for their thoughtful comments that improved this work. This work was supported by the National Science Foundation grant #1942125.

REFERENCES

[1] Julia Angwin, ProPublica, and Hannes Grassegger. 2017. Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

[2] Carolina Are. 2020. How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist Media Studies* 20, 5 (July 2020), 741–744. https://doi.org/10.1080/14680777.2020.1783805 Publisher: Routledge _eprint: https://doi.org/10.1080/14680777.2020.1783805.

[3] Carolina Are. 2021. The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies* 0, 0 (May 2021), 1–18. https://doi.org/10.1080/14680777.2021.1928259 Publisher: Routledge _eprint: https://doi.org/10.1080/14680777.2021.1928259.

[4] Paul M. Barrett and J. Grant Sims. 2021. *False Accusation: The Unfounded Claim that Social Media Companies Censor Conservatives.* Technical Report. NYU Stern Center for Business and Human Rights. https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/6011e68dec2c7013d3caf3cb/1611785871154/NYU+False+Accusation+report_FINAL.pdf

[5] Jeffrey Layne Blevins, Ezra Edgerton, Don P. Jason, and James Jaehoon Lee. 2021. Shouting Into the Wind: Medical Science versus "B.S." in the Twitter Maelstrom of Politics and Misinformation About Hydroxychloroquine. *Social Media + Society* 7, 2 (April 2021), 20563051211024977. https://doi.org/10.1177/20563051211024977 Publisher: SAGE Publications Ltd.

[6] Danielle Blunt, Emily Coombes, Shanelle Mullin, and Ariel Wolf. 2020. *Posting into the Void.* Technical Report. Hacking//Hustling.

[7] Danielle Blunt and Ariel Wolf. 2020. *Erased: The Impact of FOSTA-SESTA & the Removal of Backpage.* Technical Report. Hacking//Hustling.

[8] Catalina Botero-Marino, Jamal Greene, Michael W. McConnell, and Helle Thorning-Schmidt. 2020. Opinion | We Are a New Board Overseeing Facebook. Here's What We'll Decide. *The New York Times* (May 2020). https://www.nytimes.com/2020/05/06/opinion/facebook-oversight-board.html

[9] Carolyn Bronstein. 2020. Pornography, Trans Visibility, and the Demise of Tumblr. *TSQ: Transgender Studies Quarterly* 7, 2 (May 2020), 240–254. https://doi.org/10.1215/23289252-8143407 Publisher: Duke University Press.

[10] Taina Bucher. 2018. *If ... Then: Algorithmic Power and Politics.* Oxford University Press. Google-Books-ID: 2GdaDwAAQBAJ.

[11] Robyn Caplan. 2018. *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches.* Technical Report. Data&Society. https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

[12] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 31:1–31:22. https://doi.org/10.1145/3134666

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

[14] Alexander Cheves. 2018. The Dangerous Trend of LGBTQ Censorship on the Internet. (Dec. 2018). https://www.out.com/out-exclusives/2018/12/06/dangerous-trend-lgbtq-censorship-internet

[15] Aymar Jean Christian, Faithe Day, Mark Díaz, and Chelsea Peterson-Salahuddin. 2020. Platforming Intersectionality: Networked Solidarity and the Limits of Corporate Social Media:. *Social Media + Society* (Aug. 2020). https://doi.org/10.1177/2056305120933301 Publisher: SAGE PublicationsSage UK: London, England.

[16] Danielle Keats Citron. 2014. *Hate Crimes in Cyberspace.* Harvard University Press.

[17] Danielle Keats Citron and Helen Norton. 2011. Intermediaries and Hate Speech: Fostering Digital Citizenship for our Information Age. *BUL Review* (2011), 51.

[18] Ben Collins. 2021. Increasingly militant 'Parler refugees' and anxious QAnon adherents prep for doomsday. https://www.nbcnews.com/tech/internet/increasingly-militant-parler-refugees-anxious-qanon-adherents-prep-doomsday-n1254775

[19] Kate Crawford and Tarleton Gillespie. 2014. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* (July 2014), 1–19. https://doi.org/10.1177/1461444814543163

[20] Cristina Criddle. 2020. Transgender users accuse TikTok of censorship. *BBC News* (Feb. 2020). https://www.bbc.com/news/technology-51474114

[21] Gilles Deleuze and Felix Guattari. 1987. *A Thousand Plateaus: Capitalism and Schizophrenia.* U of Minnesota Press. Google-Books-ID: C948Tsr72woC.

[22] Dolores Delgado Bernal. 1997. *Chicana school resistance and grassroots leadership: providing an alternative history of the 1968 East Los Angeles blowouts.* Ph.D. Dissertation. University of California, Los Angeles.

[23] Christina Dinar. 2021. *The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act.* Technical Report. Heinrich-Böll-Stiftung. 23 pages.

[24] Natasha Duarte, Emma Llansó, and Anna Loup. 2018. Mixed Messages? The Limits of Automated Social Media Content Analysis. In *2018 Conference on Fairness, Accountability, and Transparency.* https://cdt.org/files/2017/12/FAT-conference-draft-2018.pdf

[25] Stefanie Duguay, Jean Burgess, and Nicolas Suzor. 2018. Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. 16. https://doi.org/10.1177/1354856518781530

[26] Elizabeth Dwoskin, Nitasha Tiku, and Heather Kelly. 2020. Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show. *Washington Post* (Dec. 2020). https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/

[27] Facebook. 2021. Community Standards. https://www.facebook.com/communitystandards/

[28] Facebook. 2021. Community Standards: Adult Nudity and Sexual Activity. https://www.facebook.com/communitystandards/adult_nudity_sexual_activity

[29] Facebook. 2021. Community Standards: False News. https://www.facebook.com/communitystandards/false_news

[30] Facebook. 2021. Community Standards: Hate Speech. https://www.facebook.com/communitystandards/hate_speech

[31] Facebook. 2021. Community Standards: Violence and Incitement. https://www.facebook.com/communitystandards/credible_violence

[32] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20).* Association for Computing Machinery,

Honolulu, HI, USA, 1–14. https://doi.org/10.1145/3313831.3376293
[33] 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 (May 2020), 040:1–040:28. https://doi.org/10.1145/3392845
[34] Facebook for Media. [n.d.]. Working to Stop Misinformation and False News. https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news
[35] Electronic Frontier Foundation. 2019. EFF Project Shows How People Are Unfairly "TOSsed Out" By Platforms' Absurd Enforcement of Content Rules. https://www.eff.org/press/releases/eff-project-shows-how-people-are-unfairly-tossed-out-platforms-absurd-enforcement
[36] Electronic Frontier Foundation. 2019. What Tumblr's Ban on 'Adult Content' Actually Did. https://www.eff.org/tossedout/tumblr-ban-adult-content
[37] Paulo Freire. 1970. *Education for critical consciousness.* Continuum Publishing Company, New York, NY, USA.
[38] Paulo Freire. 1973. *Pedagogy of the oppressed.* The Seabury Press, New York, NY, USA.
[39] Ahiza Garcia-Hodges. 2021. Apple App Store, Google Play suspend Parler pending better moderation. https://www.nbcnews.com/tech/social-media/google-play-suspends-parler-until-app-develops-moderation-policies-n1253609
[40] Meira Gebel. 2020. Black Creators Say TikTok Still Secretly Hides Their Content. https://www.digitaltrends.com/social-media/black-creators-claim-tiktok-still-secretly-blocking-content/
[41] Ysabel Gerrard and Helen Thornham. 2020. Content moderation: Social media's sexist assemblages. *New Media & Society* 22, 7 (July 2020), 1266–1286. https://doi.org/10.1177/1461444820912540 Publisher: SAGE Publications.
[42] Tarleton Gillespie. 2017. Governance of and by platforms. In *The SAGE Handbook of Social Media.* SAGE, New York, 30.
[43] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press, New Haven.
[44] Henry Giroux. 1983. Theories of reproduction and resistance in the new sociology of education: a critical analysis. *Harvard Educational Review* (1983).
[45] Kayla Gogarty, Spencer Silva, Carly Evans, and Media Matters. 2020. *A new study finds that Facebook is not censoring conservatives despite their repeated attacks.* Technical Report. https://www.mediamatters.org/facebook/new-study-finds-facebook-not-censoring-conservatives-despite-their-repeated-attacks
[46] Alessandra Gomes and Dennys e Thiago Dias Oliva Antonialli. 2019. Drag queens and Artificial Intelligence: should computers decide what is 'toxic' on the internet? http://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/
[47] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17 (2015), 42–109. https://heinonline.org/HOL/P?h=hein.journals/yjolt17&i=42
[48] Jessica Guynn. 2017. Facebook apologizes to black activist who was censored for calling out racism. *USA Today* (Aug. 2017). https://www.usatoday.com/story/tech/2017/08/03/facebook-ijeoma-oluo-hate-speech/537682001/
[49] Jessica Guynn. 2019. Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech. https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/
[50] Oliver L. Haimson. 2018. Social Media as Social Transition Machinery. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 63:1–63:27. https://doi.org/10.1145/3274332
[51] Oliver L. Haimson, Jed R. Brubaker, Lynn Dombrowski, and Gillian R. Hayes. 2016. Digital Footprints and Changing Networks During Online Identity Transitions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM, New York, NY, USA, 2895–2907. https://doi.org/10.1145/2858036.2858136
[52] Oliver L. Haimson, Justin Buss, Zu Weinger, Denny L. Starks, Dykee Gorrell, and Briar Sweetbriar Baron. 2020. Trans Time: Safety, Privacy, and Content Warnings on a Transgender-Specific Social Media Site. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 124:1–124:27. https://doi.org/10.1145/3415195
[53] Oliver L. Haimson, Avery Dame-Griff, Elias Capello, and Zahari Richter. 2019. Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist Media Studies* (Oct. 2019), 1–17. https://doi.org/10.1080/14680777.2019.1678505
[54] Oliver L. Haimson and Anna Lauren Hoffmann. 2016. Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. *First Monday* 21, 6 (June 2016). https://doi.org/10.5210/fm.v21i6.6791
[55] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599. https://doi.org/10.2307/3178066 Publisher: Feminist Studies, Inc.
[56] Sandra G. Harding. 2004. *The Feminist Standpoint Theory Reader: Intellectual and Political Controversies.* Psychology Press. Google-Books-ID: qmSySHvIy5IC.
[57] Drew Harwell and Craig Timberg. 2019. Pro-Trump message board 'quarantined' by Reddit following violent threats. *Washington Post* (June 2019). https://www.washingtonpost.com/technology/2019/06/26/pro-trump-message-board-

quarantined-by-reddit-following-violent-threats/
[58] Sally Hines. 2019. The feminist frontier: on trans and feminism. *Journal of Gender Studies* 28, 2 (Feb. 2019), 145–157. https://doi.org/10.1080/09589236.2017.1411791
[59] Anna Lauren Hoffmann and Anne Jonas. 2017. Recasting Justice for Internet and Online Industry Research Ethics. In *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts*, Michael Zimmer and Katharina Kinder-Kurlanda (Eds.). Peter Lang Publishing, 3–19.
[60] Twitter Inc. 2021. Permanent suspension of @realDonaldTrump. https://blog.twitter.com/en_us/topics/company/2020/suspension.html
[61] Instagram. 2021. Combatting Misinformation on Instagram. https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram
[62] Instagram. 2021. Community Guidelines | Instagram Help Center. https://help.instagram.com/477434105621119?ref=ig_about
[63] Mike Isaac. 2020. Reddit, Acting Against Hate Speech, Bans 'The_Donald' Subreddit. *The New York Times* (June 2020). https://www.nytimes.com/2020/06/29/technology/reddit-hate-speech.html
[64] Mike Isaac and Kellen Browning. 2020. Fact-Checked on Facebook and Twitter, Conservatives Switch Their Apps. *The New York Times* (Nov. 2020). https://www.nytimes.com/2020/11/11/technology/parler-rumble-newsmax.html
[65] Tracy Jan and Elizabeth Dwoskin. 2017. A white man called her kids the n-word. Facebook stopped her from sharing it. *Washington Post* (July 2017). https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html
[66] 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 (Nov. 2019), 192:1–192:33. https://doi.org/10.1145/3359294
[67] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2 (March 2018), 12:1–12:33. https://doi.org/10.1145/3185593
[68] Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2019. Bias Misperceived:The Role of Partisanship and Misinformation in YouTube Comment Moderation. *Proceedings of the International AAAI Conference on Web and Social Media* 13 (July 2019), 278–289. https://ojs.aaai.org/index.php/ICWSM/article/view/3229
[69] Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2020. Reasoning about Political Bias in Content Moderation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (April 2020), 13669–13672. https://doi.org/10.1609/aaai.v34i09.7117
[70] Jeffrey M. Jones. 2021. LGBT Identification Rises to 5.6% in Latest U.S. Estimate. https://news.gallup.com/poll/329708/lgbt-identification-rises-latest-estimate.aspx Section: Politics.
[71] Ben Kew. 2018. Poll: Two-Thirds of Conservatives Don't Trust Facebook, Believe Social Media Censors Conservatives. https://www.breitbart.com/tech/2018/08/29/poll-two-thirds-of-conservatives-dont-trust-facebook-believe-social-media-censors-conservatives/
[72] Kate Klonick. 2020. The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. *the yale law journal* (2020), 82.
[73] Kate Klonick. 2021. Inside the Making of Facebook's Supreme Court. https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court
[74] Karen Kornbluh, Adrienne Goldstein, and Eli Weiner. 2020. *New Study by Digital New Deal Finds Engagement with Deceptive Outlets Higher on Facebook Today Than Run-up to 2016 Election.* Technical Report. The German Marshall Fund of the United States. https://www.gmfus.org/blog/2020/10/12/new-study-digital-new-deal-finds-engagement-deceptive-outlets-higher-facebook-today
[75] Jessa Lingel. 2017. *Digital Countercultures and the Struggle for Community* (1 edition ed.). The MIT Press, Cambridge, MA.
[76] Jessa Lingel. 2019. The gentrification of the internet. https://culturedigitally.org/2019/03/the-gentrification-of-the-internet/
[77] Dottie Lux and Lil Miss Hot Mess. 2017. Facebook's Hate Speech Policies Censor Marginalized Users. *Wired* (Aug. 2017). https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/
[78] Brandeis Marshall. 2021. *Algorithmic misogynoir in content moderation practice.* Technical Report. Heinrich-Böll-Stiftung. 17 pages.
[79] Nataliez Martinez and Media Matters. 2018. *Study: Analysis of top Facebook pages covering American political news.* Technical Report. https://www.mediamatters.org/facebook/study-analysis-top-facebook-pages-covering-american-political-news
[80] Natalie Martinez and Media Matters. 2019. *Study: Facebook is still not censoring conservatives.* Technical Report. https://www.mediamatters.org/facebook/study-facebook-still-not-censoring-conservatives

[81] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (March 2017), 329–346. https://doi.org/10.1177/1461444815608807
[82] Peter McLaren. 1994. *Life in schools: an introduction to critical pedagogy in the foundations of education* (2nd ed.). Longman, New York, NY, USA.
[83] Danaë Metaxa, Joon Sung Park, James A. Landay, and Jeff Hancock. 2019. Search Media and Elections: A Longitudinal Investigation of Political Search Results. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 129:1–129:17. https://doi.org/10.1145/3359231
[84] Deirdre K. Mulligan, Daniel Kluttz, and Nitin Kohli. 2019. *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*. SSRN Scholarly Paper ID 3311894. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3311894
[85] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (Nov. 2018), 4366–4383. https://doi.org/10.1177/1461444818773059
[86] Lisa Nakamura. 2014. 'I WILL DO EVERYthing That Am Asked': Scambaiting, Digital Show-Space, and the Racial Violence of Social Media. *Journal of Visual Culture* 13, 3 (Dec. 2014), 257–274. https://doi.org/10.1177/1470412914546845 Publisher: SAGE Publications.
[87] Viviane Namaste. 2000. *Invisible Lives: The Erasure of Transsexual and Transgendered People*. University of Chicago Press. Google-Books-ID: Pq5jwRVbvY8C.
[88] Casey Newton. 2019. The real bias on social networks isn't against conservatives. https://www.theverge.com/interface/2019/4/11/18305407/social-network-conservative-bias-twitter-facebook-ted-cruz
[89] Casey Newton. 2021. Conservative social networks keep making the same mistake. https://www.platformer.news/p/conservative-social-networks-keep
[90] Ihudiya Finda Ogbonnaya-Ogburu, Angela D. R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems*. 16.
[91] onlinecensorship.org. [n.d.]. onlinecensorship.org — Submit Your Report. https://onlinecensorship.org/takedowns/new
[92] onlinecensorship.org. 2018. Offline-Online. https://onlinecensorship.org/content/infographics
[93] John E Pachankis. 2007. The psychological implications of concealing a stigma: A cognitive-affective-behavioral model. *Psychological Bulletin* 133, 2 (March 2007), 18. https://doi.org/10.1037/0033-2909.133.2.328
[94] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 12. https://doi.org/10.1145/2858036.2858389
[95] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work - GROUP '16*. ACM Press, Sanibel Island, Florida, USA, 369–374. https://doi.org/10.1145/2957276.2957297
[96] Delroy L. Paulhus. 1991. *Measurement and Control of Response Bias*.
[97] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Robert West. 2020. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *arXiv:2010.10397 [cs]* (Oct. 2020). http://arxiv.org/abs/2010.10397 arXiv: 2010.10397.
[98] Sarah T Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. In *Intersectional Internet: Race, Sex, Class and Culture Online*. Peter Lang, 12.
[99] Sarah T. Roberts. 2018. Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday* 23, 3 (March 2018). https://doi.org/10.5210/fm.v23i3.8283
[100] Kevin Roose. 2020. The President Versus the Mods. *The New York Times* (May 2020). https://www.nytimes.com/2020/05/29/technology/trump-twitter.html
[101] Mey Rude. 2019. Trace Lysette Is Latest Trans Woman Banned By Tinder. https://www.out.com/transgender/2019/9/19/trace-lysette-latest-trans-woman-be-banned-tinder Library Catalog: www.out.com.
[102] Salty. 2019. Exclusive: An Investigation into Algorithmic Bias in Content Policing on Instagram. https://www.saltyworld.net/algorithmicbiasreport-2/
[103] Salty. 2020. Shadowbanning is a Thing — and It's Hurting Trans and Disabled Advocates. https://saltyworld.net/shadowbanning-is-a-thing-and-its-hurting-trans-and-disabled-advocates/ Library Catalog: saltyworld.net Section: Algorithmic Bias.
[104] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 155:1–155:27. https://doi.org/10.1145/3274424
[105] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. *New Media & Society* (March 2020), 1461444820913122. https://doi.org/10.1177/1461444820913122 Publisher: SAGE Publications.

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

[107] Jack Sidnell. 2002. *Outline of African American Vernacular English (AAVE) Grammar.* Technical Report. https://cdt.org/wp-content/uploads/2017/11/Outline_of_AAVE_grammar___Jack_Sidnell_2002_1_Afr.pdf

[108] Olivia Solon. 2020. Facebook ignored racial bias research, employees say. https://www.nbcnews.com/tech/tech-news/facebook-management-ignored-internal-research-showing-racial-bias-current-former-n1234746

[109] Daniel Solórzano and Dolores Delgado Bernal. 2001. Critical race theory, transformational resistance and social justice: Chicana and Chicano students in an urban context. *Urban Education* (2001).

[110] Clare Southerton, Daniel Marshall, Peter Aggleton, Mary Lou Rasmussen, and Rob Cover. 2020. Restricted modes: Social media, content classification and LGBTQ sexual citizenship. *New Media & Society* (Feb. 2020), 1461444820904362. https://doi.org/10.1177/1461444820904362 Publisher: SAGE Publications.

[111] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. 2019. How to Do Better with Gender on Surveys: A Guide for HCI Researchers. *Interactions* (Aug. 2019), 4.

[112] Sanna Spišák, Elina Pirjatanniemi, Tommi Paalanen, Susanna Paasonen, and Maria Vihlman. 2021. Social Networking Sites' Gag Order: Commercial Content Moderation's Adverse Implications for Fundamental Sexual Rights and Wellbeing. *Social Media + Society* 7, 2 (April 2021), 20563051211024962. https://doi.org/10.1177/20563051211024962 Publisher: SAGE Publications Ltd.

[113] Liam Stack. 2019. Trump Wants Your Tales of Social Media Censorship. And Your Contact Info. *The New York Times* (May 2019). https://www.nytimes.com/2019/05/15/us/donald-trump-twitter-facebook-youtube.html

[114] Anselm Strauss and Juliet M. Corbin. 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory.* SAGE Publications. Google-Books-ID: tBcEjwEACAAJ.

[115] Nicolas P. Suzor. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives.* Cambridge University Press. Google-Books-ID: EjGdDwAAQBAJ.

[116] Nicolas P Suzor. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13 (2019), 1526–1543.

[117] Jeanna Sybert. 2021. The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *New Media & Society* (Feb. 2021), 1461444821996715. https://doi.org/10.1177/1461444821996715 Publisher: SAGE Publications.

[118] Twitter. 2021. COVID-19 misleading information policy. https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

[119] Twitter. 2021. The Twitter rules: safety, privacy, authenticity, and more. https://help.twitter.com/en/rules-and-policies/twitter-rules

[120] Twitter. 2021. Twitter's policy on hateful conduct | Twitter Help. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[121] Twitter. 2021. Twitter's sensitive media policy | Twitter Help. https://help.twitter.com/en/rules-and-policies/media-policy

[122] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact* 4, CSCW2 (2020), 22.

[123] Emily A. Vogels, Andrew Perrin, and Monica Anderson. 2020. *Most Americans Think Social Media Sites Censor Political Viewpoints.* Technical Report. Pew Research Center. https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/

[124] Cristan Williams. 2016. Radical Inclusion: Recounting the Trans Inclusive History of Radical Feminism. *TSQ: Transgender Studies Quarterly* 3, 1-2 (May 2016), 254–258. https://doi.org/10.1215/23289252-3334463

[125] Jillian C. York. 2021. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism.* Verso Books. Google-Books-ID: SNwfEAAAQBAJ.

[126] Tara J. Yosso. 2005. Whose culture has capital? A critical race theory discussion of community cultural wealth. 8 (March 2005), 24. https://doi.org/10.1080/1361332052000341006

[127] Savvas Zannettou. 2021. "I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. *arXiv:2101.07183 [cs]* (Jan. 2021). http://arxiv.org/abs/2101.07183 arXiv: 2101.07183.

[128] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: Building Governance in Online Communities. In *Proceedings of UIST*. 14.

[129] Mark Zuckerberg. 2021. Mark Zuckerberg announces Trump banned from Facebook indefinitely. https://www.facebook.com/zuck/posts/10112681480907401

A APPENDIX: SURVEY INSTRUMENT

This appendix only includes parts of the survey that were included in this paper's analysis. Several questions were adapted from OnlineCensorship.org [91] and Myers West [85].

Survey 1 only included sections A.1 and A.4.
Survey 2 included all sections A.1 - A.4.

A.1 Screening and basic questions

(1) Do you live in the United States? [Yes; No]
(2) How old are you? [Younger than 18; 18-24; 25-34; 35-44; 45-54; 55-64; 65 or older]
(3) Within the last year, have you had content taken down from a social media site for reasons you disagreed with? [Yes; No]
(4) Within the last year, has your account been taken down from a social media site for reasons you disagreed with? [Yes; No]

A.2 Questions about content removals (only asked if participants specified they experienced a content removal in the past year)

(5) Please describe the most memorable *content* takedown that you experienced in the past year. What content was removed? [open-ended]
(6) On which social media platform(s) did the content takedown occur? [Discord; Facebook; Instagram; LinkedIn; Pinterest; Quora; Reddit; SnapChat; TikTok; Tumblr; Twitter; WhatsApp; YouTube; Other:_____]
(7) Was the content taken down by the platform itself, or by a social media user in a moderator/admin role? [The platform removed my content; A social media user in a moderator/admin role (i.e., not employed by the platform) removed my content; I'm not sure]
(8) What reason was given for the content takedown? [open-ended]
(9) Why do you think the content takedown happened? [open-ended]
(10) How, if at all, did the content takedown impact you personally? [open-ended]
(11) Did you appeal the content takedown decision? [Yes; No]
(12) What was the outcome of the appeal? [open-ended]
(13) Are there any further details about your experience you would like to provide? [open-ended]

A.3 Questions about account removals (only asked if participants specified they experienced an account removal in the past year)

(14) Please describe the most memorable *account* takedown that you experienced in the past year. What account was taken down?
(15) On which social media platform(s) did the account takedown occur? [Discord; Facebook; Instagram; LinkedIn; Pinterest; Quora; Reddit; SnapChat; TikTok; Tumblr; Twitter; WhatsApp; YouTube; Other:_____]
(16) Was your account taken down by the platform itself, or by a social media user in a moderator/admin role? [The platform removed my content; A social media user in a moderator/admin role (i.e., not employed by the platform) removed my content; I'm not sure]
(17) What reason was given for the account takedown? [open-ended]
(18) Why do you think the account takedown happened? [open-ended]
(19) How, if at all, did the account takedown impact you personally? [open-ended]
(20) Did you appeal the account takedown decision? [Yes; No]
(21) What was the outcome of the appeal? [open-ended]
(22) Are there any further details about your experience you would like to provide? [open-ended]

A.4 Demographic questions

(23) What is your gender? (select all that apply) [Woman; Man; Non-binary; Prefer not to disclose; Prefer to self-describe:_____] (as recommended by [111])
(24) Are you transgender? [Yes; No; Prefer not to disclose]
(25) Choose one or more races/ethnicities that you consider yourself to be. (select all that apply) [American Indian or Alaska Native; Asian; Black or African American; Hispanic or Latino; Middle Eastern; Native Hawaiian or Pacific Islander; White]
(26) What is your sexual orientation? [Straight; Gay; Lesbian; Bisexual; Pansexual; Queer; Asexual; Prefer not to disclose; Prefer to self-describe:_____]
(27) Information about income is very important to understand. Would you please give your best guess? Please indicate the answer that includes your entire household income in 2019 before taxes. [Less than $20,000; $20,000 to $34,999; $35,000 to $49,999; $50,000 to $74,999; $75,000 to $99,999; $100,000 to $149,999; $150,000 or more]
(28) What is the highest level of school you have completed or the highest degree you have received? [Less than high school; High school or equivalent; Some college or two-year associate's degree; Bachelor's degree; Some graduate school; Master's or professional degree; Doctoral degree]
(29) In general, how would you describe your political views? [Very conservative; Conservative; Moderate; Liberal; Very liberal]

B APPENDIX: RELEVANT SOCIAL MEDIA SITE POLICIES

Table 8. Relevant social media site policies as of April 2021, for the three sites most frequently used by participants. While the full text of each policy is too long to include here, we include some relevant excerpts.

Code	Excerpt from site policy wording
offensive/inappropriate content	
Facebook	"*We are committed to making Facebook a safe place. Expression that threatens people has the potential to intimidate, exclude or silence others and isn't allowed on Facebook.*" [27] "*We remove content, disable accounts, and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety. We also try to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety.*" [31]
Twitter	"*You may not post media that is excessively gory or share violent or adult content within live video or in profile or header images.*" [119]
Instagram	"*We understand that many people use Instagram to share important and newsworthy events. Some of these issues can involve graphic images. Because so many different people and age groups use Instagram, we may remove videos of intense, graphic violence to make sure Instagram stays appropriate for everyone.*" [62]
adult content	
Facebook	"*Our nudity policies have become more nuanced over time. We understand that nudity can be shared for a variety of reasons, including as a form of protest, to raise awareness about a cause, or for educational or medical reasons. Where such intent is clear, we make allowances for the content.*" [28]
Twitter	"*Media depicting sexual violence and/or assault is also not permitted.*" [121]
Instagram	"*We know that there are times when people might want to share nude images that are artistic or creative in nature, but for a variety of reasons, we don't allow nudity on Instagram.*" [62]
misinformation	
Facebook	"*Do not post: Misinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm. Additionally, we have specific rules and guidance regarding content related to COVID-19 and vaccines.*" [29]
Twitter	"*You may not use Twitter's services to share false or misleading information about COVID-19 which may lead to harm.* " [118]
Instagram	"*When content has been rated as false or partly false by a third-party fact-checker, we reduce its distribution by removing it from Explore and hashtag pages, and reducing its visibility in Feed and Stories.*" [61]
content criticizing group	
Facebook	"*Do not post: Self-admission to intolerance on the basis of a protected characteristics, including but not limited to: homophobic, Islamophobic, racist*" [30]
Twitter	"*We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these [identity] categories.*" [120]
Instagram	"*It's never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases.*" [62]
hate speech	
Facebook	"*We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That's why we don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence.*" [30]
Twitter	"*You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.*" [119]
Instagram	"*We remove content that contains credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages.*" [62]

Received April 2021; revised July 2021; accepted July 2021