Content Moderation Folk Theories And Perceptions of Platform Spirit Among Marginalized Social Media Users

SAMUEL MAYWORM, University of Michigan, USA
MICHAEL ANN DEVITO, Northeastern University, USA
DAN DELMONACO, University of Michigan, USA
HIBBY THACH, University of Michigan, USA
OLIVER L. HAIMSON, University of Michigan, USA

Social media users create folk theories to help explain how elements of social media operate. Marginalized social media users face disproportionate content moderation and removal on social media platforms. We conducted a qualitative interview study (n = 24) to understand how marginalized social media users may create folk theories in response to content moderation and their perceptions of platforms' spirit, and how these theories may relate to their marginalized identities. We found that marginalized social media users develop folk theories informed by their perceptions of platforms' spirit to explain instances where their content was moderated in ways that violate their perceptions of how content moderation should work in practice. These folk theories typically address content being removed despite not violating community guidelines, along with bias against marginalized users embedded in guidelines. We provide implications for platforms, such as using marginalized users' folk theories as tools to identify elements of platform moderation systems that function incorrectly and disproportionately impact marginalized users.

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

Additional Key Words and Phrases: algorithm, algorithmic content moderation, content moderation, folk theories, platform spirit, social media, marginalization, marginalized identity

1 INTRODUCTION

Marginalized social media users (defined in our paper as social media users who experience systemic exclusion, discrimination, and social inequities based on factors such as race, ethnicity, gender identity, sexuality, disability, etc., and the intersections of these factors) often use social media for purposes unique to their communities and identities, including expressing frustration with bigotry, finding information resources unique to their community, or seeking support, connections, and solidarity from other members of their community [37, 40, 49, 51, 54, 55, 66, 67]. But compared to more privileged social groups, marginalized social media users are disproportionately targeted by social media content moderation systems for content takedowns and account bans, even when their content does not violate platform guidelines. This results in social media platforms becoming increasingly hostile environments for marginalized social media users, who can find themselves unable to use social media in the ways they find useful or enjoyable for fear of their posts and accounts being incorrectly removed [37]. Marginalized social media users and communities across a range of identities are deeply impacted

Authors' addresses: Samuel Mayworm, mayworms@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Michael Ann DeVito, m.devito@northeastern.edu, Northeastern University, Boston, Massachusetts, USA; Dan Delmonaco, delmonac@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Hibby Thach, hibby@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Oliver L. Haimson, haimson@umich.edu, University of Michigan, Ann Arbor, Michigan, USA.

ACM 2469-7818/2023/12-ART
https://doi.org/10.1145/3632741

by this hostility; for example, transgender and nonbinary users often face disproportionate removals of content related to trans-specific healthcare needs, while Black users often experience disproportionate removals of their content discussing racial justice and their own experiences with antiblackness [37]. Not only are marginalized users harmed by the disproportionate removals of their content, but they are also harmed by facing inequitable treatment on a platform in the first place, reproducing similar social and infrastructural inequities that they experience elsewhere.

Social media users often create and employ folk theories[1] to explain their experiences with unclear content and account removals, responding to content moderation practices that are invisible, difficult to understand, or seemingly contradict platform's guidelines [11, 13, 21, 46]. Folk theories can help users identify potential causes for their content and account removals [12, 21]. Users' theories can also guide their overall perceptions of a platform, their behavioral decision-making on a platform, and even whether they'll choose to continue using a platform into the future [11, 12].

When social media platforms' systems are opaque to ordinary users, users draw from a range of endogenous and exogenous information to fuel their folk theories about those platforms and their systems [14]; in the case of platforms' content moderation systems, endogenous information could come from users' personal experiences having their content and accounts moderated on platforms, while exogenous information could come in the form of articles and discussions about platforms' moderation systems [14]. One factor that can influence users' perceptions of platforms includes platform spirit, defined by DeVito as the user's own understanding of what a platform is supposed to do and be for, gathered from the platform's public statements, actions in practice, functionality, and demonstrated values, as framed by the user's own use case for the platform [11, 12]. User perceptions of platform spirit are built over time, and can take on a negative or positive valence depending on how well the platform fulfills what the user perceives as its spirit [11]. Past work has described how social media users draw from their perceptions of platforms' spirit to decide whether it is worth the effort to adapt their own behavior in order to remain on platforms [14]. Users' overall perceptions of platform spirit differ from more detailed, specific folk theories about how platforms and their systems (such as content moderation systems) work; past work has explored how marginalized users' folk theories about platforms can inform their overall perceptions of said platforms' spirit, particularly if the users' theories lead them to perceive the platform as having a negative spirit [12].

Compared to social media users generally, marginalized users are particularly likely to draw from perceptions of platforms' spirit or develop folk theories to determine how to safely use a platform whose content moderation practices disproportionately remove marginalized users' content, suppress topics related to marginalized identities, or otherwise expose marginalized users to harm [12, 37]. Because of the deeply negative and painful state of social media platforms for marginalized users, their folk theories and perceptions of platforms' spirit are particularly high-stakes, as they are especially likely to exist in response to identity-based harm. Prior work has presented the use of platform spirit and folk theories as sensemaking and behavioral guidance tools by marginalized users that help them determine how to successfully navigate these disproportionate risks in order to use social media platforms safely, or whether to use those platforms at all [12].

In this paper, to extend prior work focused on folk theorization, platform spirit, content moderation, and marginalized social media users, we examine folk theories and perceptions of platforms' spirit that marginalized social media users develop about social media moderation and content removals. We ask:

[1]Folk theories are defined as a "person's intuitive, causal explanation about a system" [26], sometimes described as less specific "systems of belief" than mental models [29], that are developed by ordinary, "non-experts" in a given field, guiding their decisions and behaviors relating to that system [28].

- **RQ1:** How do marginalized users' perceptions of social media platforms and their content moderation practices, including users' perceptions of platforms' spirit, influence users' folk theorization processes on social media platforms?
- **RQ2:** How do marginalized social media users adjust their behavior on social media platforms in response to the folk theories they develop about platforms and their content moderation practices?

To address these research questions, we interviewed 24 marginalized social media users who experienced content or account removals from social media platforms within the past year. We asked about their experiences on the platform before, during, and after the content moderation experience, and whether and how their removal may relate to their marginalized identity. We found that marginalized social media users develop folk theories not only to explain individual instances of content and account removals, but as tools that clarify the mechanics behind a social media platform's inequitable treatment of its marginalized user communities. We found that marginalized users' folk theories typically originate from initial negative identity-related moderation experiences on platforms that lead the user to perceive the platform as having a negative platform spirit, highlighting a two-way relationship between marginalized users' perceptions of platforms' spirit and their folk theories about platforms, extending DeVito's past work that described a more unidirectional relationship between marginalized users' folk theories about platforms and their perceptions (particularly negative perceptions) of those platforms' spirit [12]. We also found that users' folk theories reinforce their overall negative perception of the platform after they've been developed. Users responded to their theories and negative platform perceptions in different ways, such as adjusting their behavior and language use to avoid incorrect algorithmic moderation, reducing their use of the platform, or even leaving the platform entirely. In this study, we argue that marginalized users' social media folk theories are valuable tools that can pinpoint and identify weaknesses in social media platform design that result in inequitable social media experiences for marginalized users.

Past work has explored reasons that social media users create folk theories addressing the content and account moderation practices that they encounter on social media [14, 21, 26]. Researchers have also explored how specific marginalized user communities, such as LGBTQIA+ users [15, 37, 46], Black users [37], and disabled users [63], have been impacted by algorithmic content moderation and visibility systems, and how certain marginalized content creators uniquely utilize their folk theories to navigate their algorithmic visibility on social media platforms [12]. We expand on this by closely examining the relationship between marginalized users' negative perceptions of platforms' spirit (rooted in their experiences having content or accounts removed on their platforms) and how they develop folk theories in response to those perceptions. We also examine how users' negative perceptions of platforms' spirit can contribute to their decision to use folk theories to guide their behaviors and decision-making on platforms – including whether they decide to leave their platforms. By better understanding marginalized users' use of folk theories to identify and navigate the threats posed to them on social media platforms, platforms themselves may be able to use those theories as valuable troubleshooting tools to identify the platform design decisions that enable those threats and drive marginalized users' negative perceptions of their platforms. Understanding marginalized users' folk theories and taking them seriously can allow social media platforms to more efficiently repair the structural elements of their design that disproportionately harm their marginalized user communities.

This work makes the following contributions to the social computing literature: 1) An understanding of how marginalized social media users' disproportionate experiences with identity-related content removals inform their negative perceptions of social media platforms' spirit, which in turn inform their folk theories about platforms; 2) A description of how marginalized social media users develop folk theories to identify how platforms' tools and affordances may not work as intended, exposing marginalized users to harm; 3) An understanding of how marginalized users respond to their folk theories about social media platforms; 4) Suggestions for how

platforms can respond to and address folk theories without downplaying the real impacts disproportionate content moderation can have on marginalized users.

2 RELATED WORK

2.1 Social Media Content Moderation

Grimmelmann defines moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [36]. Content moderation is employed by social media platforms to combat abusive user behavior while removing content that is illegal or considered harmful to others [27, 32, 43, 53, 65]. Content moderation can include human moderation [58] and algorithmic content moderation, defined by Gorwa as "systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome" [35]. Dominant social media platforms have become particularly reliant on algorithmic content moderation systems due to their large volumes of users and user activity [33, 50].

Research on user perceptions of content moderation has shown that social media users often perceive content moderation as an inconsistent, frustrating, and unfair process [32, 62, 68, 72], which can result in user distrust and resentment of content moderation and of specific moderation tools [58]. This can result in users avoiding key safety tools such as reporting abusive content and users; DeVito's work demonstrates an example of this phenomenon, where transfeminine TikTok users avoided reporting transphobic content as they perceived TikTok's reporting tools as incapable of accurately recognizing transphobic abuse [12]. Previous literature has cited perceived a lack of transparency in content moderation decisions [7, 44, 64], perceived political bias [37, 44, 45, 47], instances of algorithmic moderation failing to remove abusive content [15, 38], and conflicts between community guidelines and content moderation in practice [69] as some reasons for negative user perceptions of content moderation systems. Users may also struggle to find and understand a social media platform's community guidelines, which can result in users being unaware that they've violated a platform's guidelines until their own content is moderated [69]. In this paper, we draw from past literature on content moderation and user perceptions of content moderation systems, applying these concepts in the context of marginalized social media users and their unique experiences with content moderation.

2.2 Content Moderation and Marginalized Users

Previous literature has explored some of the ways in which marginalized users, both broadly and as specific groups, experience inequitable content moderation on social media [16, 25, 34, 37, 66]. Researchers have found that marginalized social media users are disproportionately likely to have their content suppressed or removed across different platforms, with groups such as Black users, individuals with mental illness, LGBTQIA+ users, transgender users, and women, particularly women of color, being especially susceptible to content suppression and removal [10, 16, 18–20, 24, 30, 37, 52, 59, 66, 76]. Many reports focus on how specific user groups experience inequitable content moderation on specific platforms, such as drag queens experiencing disproportionate content removals on Twitter [15] or Black users having their posts discussing anti-blackness removed from Facebook [2] and TikTok [20]. Though content moderation is an important function that ensures social media platforms remove content that is illegal or potentially harmful to its users [32], the way it functions across many platforms results in further discrimination against marginalized user groups [16, 34, 37].

Marginalized users often feel frustrated by their experiences with content moderation and removal, especially if a user believes their content removal was unfair or did not violate the platform's community guidelines in the first place [37, 66]. Algorithmic content removal can be particularly frustrating to ordinary users who do not know how algorithmic moderation tools work in the first place, let alone why those tools flagged and removed their content [6, 57]. This frustration can also be exacerbated if the user finds the platform's content moderation appeal processes to be insufficient or unhelpful in addressing incorrect content removal [37]. Several major

platforms, such as Instagram [3], YouTube [1, 46] and TikTok [4, 9, 31, 46] have been criticized for algorithmically suppressing content posted by LGBTQIA+ users and BIPOC (Black, Indigenous, and People of Color) users in the past. The disproportionate removal of marginalized users' online content can also directly impact offline activism by reducing the ability for marginalized users to organize offline activity on the internet; an example includes Black Lives Matter activists whose accounts were incorrectly removed from Facebook after sharing their experiences with antiblackness, limiting their ability to communicate with their audience and organize activist activity [48, 73] Instances like those described above prevent marginalized social media users from using these platforms as freely as non-marginalized users, severely limiting their ability to speak about their experiences, seek out community support, and take advantage of other features of social media that are crucial to marginalized people. This may lead marginalized users to not trust social media platforms to moderate their content fairly or to protect them from harm [41, 66].

This study contributes to the literature by interviewing a broad range of marginalized social media users about their experiences with content moderation and removal on a variety of social media platforms, and what kind of theories these users hold about how social media content moderation works and how marginalized social media users could adjust their behavior in order to avoid excessive content moderation on their platforms. This study builds off past work related to marginalized users' content moderation experiences [12, 37] by specifically exploring how marginalized users broadly draw from their content moderation experiences to inform their perceptions of platforms' spirit, and subsequently their folk theorization processes on social media platforms.

2.3 Content Moderation and Folk Theories

Social media folk theories are users' self-created, socially informed explanations for how a social media platform's functions work in practice, and how this affects the platform's user base [14, 21, 26, 75]. Content moderation folk theories can address several elements of social media content moderation, such as the platform's community guidelines or its algorithmic moderation tools. Users create "non-professional perspectives via first-hand experience and social interactions" [21] on how their social media platforms work, which includes content moderation practices. Examples could include user theories about whether their content is being flagged by other users for removal [57], or users creating their own explanations for content moderation when they feel a social media platform did not sufficiently explain why their content was removed [42]. Social media folk theories are sometimes inaccurate or limited as informational tools as they are created by ordinary users without system-level knowledge of how social media platforms function, particularly regarding social media algorithmic systems [17, 26]. A users' social media folk theories are rarely concrete either, but are "developed, tested, and re-evaluated" through their continued use of the social media platform [11, 13, 26]. Still, existing literature recognizes social media folk theories, including inaccurate or highly malleable ones, as a "valuable domain for research" that offer insight on user perceptions of and experiences on social media platforms [74] that are otherwise "black-boxed" and invisible to observers [14].

Previous research has explored how general social media users [17, 21] and marginalized users [13, 46] create folk theories to explain their experiences on social media, including their negative interactions with content moderation and removals [57], and to guide their behaviors on social media platforms. Previous research has also described marginalized social media users' behavioral responses to folk theories about specific platforms, such as marginalized TikTok users theorizing about and algorithmically resisting the suppression of identity-related content on the platform [46]. This study contributes to the literature by exploring social media folk theories specifically marginalized social media users develop across a range of social media platforms in response to content moderation and removal, giving insight into the unique obstacles faced by marginalized users on social media and their inequitable relationship with content moderation compared to non-marginalized users. This study explores the specific ways marginalized social media users' folk theories guide their behaviors and activity

on social media platforms, and how these theories contribute to their overall perception of their position and belonging on those platforms.

3 METHODS

To answer our research questions, we conducted an interview study with n = 24 participants. This research was reviewed and deemed exempt from oversight by our university's Institutional Review Board (IRB)[2].

3.1 Participant Recruitment

We recruited participants in three ways. First, we contacted participants from our prior survey study who indicated that they would like to participate in a follow-up interview (n = 6). Next, we promoted the study via our social media accounts on Twitter and invited interested people to fill out a screening survey; our post was retweeted many times and reached far beyond our personal networks. Based on results of the screening survey, we contacted interested participants for interviews (n = 6). Finally, we used a research recruiting service and its internal screening survey process (n = 12). We screened for adult social media users from marginalized groups (i.e., racial/ethnic minorities, gender and sexual minorities) who stated that their content or accounts were removed from a social media platform in the past year for reasons they disagreed with. To ensure that our sample was diverse and included people from marginalized groups, the screening surveys asked participants for their age, gender, race/ethnicity, LGBTQIA+ status, and whether they specifically are transgender, nonbinary, or both. We used open text in our recruiting surveys for gender and sexuality in order to respect and capture the diversity of terminology and self-identification within the queer and trans population. We present a roughly classified overview here for the reader's convenience. The participants in this study represented a range of genders: 42% of participants were cisgender women, 29% were nonbinary individuals, 21% were cisgender men, 4% were transgender women, and 4% were transgender men. The participants in this study represented a range of sexualities: 63% of participants reported identifying as either gay, lesbian, bisexual, pansexual, aromantic, or queer. The participants in this study represented a range of racial/ethnic backgrounds, with 92% of participants being racial/ethnic minorities; 38% of participants were Asian, 23% were Black, 23% were Latino, Latina, or Latinx, 8% were White, 4% were Middle Eastern, 4% were Native Hawaiian/Pacific Islander, and 4% were mixed race. The participants in this study ranged in age from their early 20s to their mid-40s, with a mean age of 28 and a standard deviation of 6.68.

3.2 Data Collection

We contacted 26 people from the pool of eligible participants to schedule an interview date and time. Of these, 24 ultimately participated in their scheduled interview. 23 interviews were conducted remotely over Zoom and recorded for audio transcription, and 1 interview with a deaf participant was conducted through text over email. Interviews lasted on average 51.65 minutes (sd = 11.06 minutes, range: 38 - 84 minutes). Participants completed the informed consent process before proceeding with the interview. Interviews were conducted by the first, second, and third authors. Multiple interviewers were typically present during interviews; when multiple interviewers were present, one took notes while the other conducted the majority of the interview itself. The interview presented a series of questions about participants' content or account removals, asking them to describe the removals, whether they thought the removals were incorrect, and how the removal experience may have

[2]At our institution, interview studies are generally deemed exempt from IRB oversight; IRB oversight is usually reserved for medical trials and studies with more in-depth or long-term interactions with participants. However, we took substantial precautions to practice ethical research and ensure that we protected participants' data, such as giving all participants anonymized participant numbers for audio recording and reporting purposes, restricting interview data access to the research team and transcribers bound to a confidentiality agreement, storing data on the research team's secure password-protected computers and secure servers, and deleting all interview audio recordings once transcripts were created and verified.

related to their marginalized identities. They were asked further questions about their perceptions of content moderation and community guidelines on the platforms they use, and how content moderation and community guidelines on platforms could be improved for marginalized users. Participants received $30 for participating in the interview study.

3.3 Data Analysis

All interviews were audio recorded, auto-transcribed through the Otter.ai text transcription service, and manually corrected by the authors of the paper and two other members of the research team. Three authors conducted open coding [5] and direct coding [8] in Atlas.ti. First, these three authors coded the same interview transcript separately, and two authors coded an additional interview transcript separately. The research team then met to discuss codes and collaboratively refine the codebook. We then coded the remaining transcripts individually; we used the codebook to code the remaining transcripts, and collaboratively updated it with new codes developed throughout subsequent analysis. We met frequently throughout the data analysis process to discuss the codebook, new codes and themes, and how we were applying codes to data. We reached theoretical saturation (defined by Corbin & Strauss as "the point in category development at which no new properties, dimensions, or relationships emerge during analysis") before the open coding of all the transcripts was complete, and transitioned to direct coding [8] of the remaining transcripts targeted at clear themes and concepts from the codebook. The authors then conducted axial coding to group codes into larger categories [8]. Themes that we developed in our data analysis include: algorithmic content moderation, community guidelines, unequal moderation of marginalized users and their content, general content moderation practices, abusive user behaviors and moderation, content visibility and suppression, the relationship between content moderation and social media platforms/corporations, and user behavior changes in response to folk theories and moderation; this paper predominantly focuses on the first three.

3.4 Positionality

The authors of this paper collectively represent a broad spectrum of marginalized identities and lived experiences; the team includes multiple authors with lived experience across queer, trans, and nonbinary identities, as well as multiple authors who represent a mixed-race background. The authors are all marginalized social media users who are familiar with (and have experienced) the various kinds of identity-based harm faced by marginalized users on the internet. The broad range of marginalized identities held by the authors benefits their collective ability to interpret and understand the experiences faced by the study participants. However, it is also important to note the limitations of the authors' backgrounds; while multiple authors are people of color, none of the authors are Black or experience antiblackness, which limits their understanding of the experiences faced by Black study participants, who represented 23% of the study participants.

4 RESULTS

Each study participant shared their perceptions of the platforms on which they experienced identity-related content removals that they considered incorrect or unfair. Participants' content removals took place across a range of highly-trafficked social media platforms, such as Facebook, Instagram, Twitter, TikTok, and Reddit. All participants stated that they perceived social media platforms as being biased against marginalized users in some way. This led participants to perceive these platforms as having a negative platform spirit, as the removals of participants' identity-related content conflicted with their perceptions of how platforms' guidelines, goals, and values should work. Participants addressed these conflicts by developing folk theories that explain the dissonance between their perception of how social media platforms should work and their experiences with identity-related content removals that informed their negative perceptions of the platforms' spirit. These theories

typically highlighted specific parts of a platform' design (such as its guidelines or algorithmic moderation tools) that marginalized users perceived as causing or enabling their negative experiences on the platform. Participants generally trusted folk theories as behavioral guides more than they trusted platforms' guidelines themselves, as many users perceived platform guidelines to be flawed, unclear, or not enforced equally between marginalized and non-marginalized users. Participants responded to their folk theories by adjusting their behavior on social media platforms; the most commonly reported behavioral change was either avoiding the use of explicitly identity-related vocabulary or using coded language as alternatives for explicitly identity-related vocabulary to avoid flagging algorithmic moderation tools. Users also reported significantly reducing their use of platforms or outright leaving them after experiencing disproportionate moderation. Platform spirit played a range of roles throughout participants' folk theorization processes. Participants' perceptions of platform spirit (guided by their experiences with content moderation and removals) served as information guiding their folk theorization; in turn, participants' negative folk theories reinforced their negative perceptions of platforms' spirit, degrading their relationships with these platforms.

In what follows, we describe participants' perceptions of platforms' spirit based on their experiences having content removed from those platforms. We then describe how participants developed folk theories about platforms and their content moderation practices in response to their negative perceptions of platforms' spirit. Afterward, we describe participants' behavioral responses guided by their folk theories, including their decision-making about how to behave on social media platforms and whether they decide to continue using those platforms at all.

4.1 Users' Perceptions of Platform Spirit

Despite social media platforms' stated goals of inclusivity and safety for a diverse community of users [23, 56, 70, 71], participants regularly experienced abuse from bigoted users, disproportionate removals of their identity-related content, and other obstacles that prevented them from using social media platforms freely or safely. These challenges drove participants to perceive their social media platforms as having a **negative platform spirit**, as the users' experiences with abuse and incorrect content removals directly contradicted the platforms' stated goals of inclusivity, safety, and freedom of self-expression for all users, violating users' understanding of what their platforms are supposed to do by failing to protect them from abuse and incorrect content moderation. For example, P3, an Asian nonbinary Instagram user who experienced transphobic harassment and incorrect removals of their selfies on the platform, shared their thoughts on Instagram's addition of a user pronoun feature:

> *[Instagram] was definitely long due for adding pronouns, but [Instagram] also doesn't do anything when people abuse or ridicule the pronoun feature. It doesn't matter if you report somebody for saying that their pronouns are 'that/bitch' or something, because Instagram will reinforce the rights of that user to 'express themselves how they want.' I've tried it!*

P3 later elaborated on how Instagram's handling of their pronoun feature damaged their overall perception of Instagram as a platform:

> *I think it was very performative of [Instagram] to include pronouns with no intention to back up their rules. It felt similar to rainbow capitalist ideas of throwing in this "confetti of representation" without giving the represented any real power to speak for their communities. And then [Instagram] does exactly what one would expect, which is not defending marginalized identities. With the use of their pronoun feature, the Instagram world was supposed to become more inclusive... but it really just opened the door to new problems.*

Though Instagram introduced a user pronoun feature that should benefit trans and nonbinary users, P3 witnessed many instances of transphobic users abusing the feature that went unmoderated on the platform. This led P3 to feel frustrated with Instagram and its perceived unwillingness to confront transphobic abuse, prioritizing the "self-expression" of its abusive users over the safety and well-being of its trans and nonbinary

users. The contradiction between Instagram's stated "support" of trans and nonbinary users and the unmoderated transphobic abuse of its pronoun feature led P3 to develop a negative perception of Instagram's platform spirit, as they now perceived Instagram's "support" of trans and nonbinary users as performative instead of genuine.

4.1.1 Platform Spirit and Algorithmic Moderation Tools. Participants held particularly negative perceptions of social media platforms' algorithmic moderation tools. These tools were generally perceived as aggressively removing or suppressing marginalized users' identity-related content, even when their content follows the platform's community guidelines. P4, a transgender man whose transition-related surgery photos were incorrectly removed for "nudity" on Facebook, shared his perception of Facebook's algorithmic moderation tools after the incorrect removal took place:

> *There's an exception [to Facebook's nudity guidelines] for transition-related surgery [images]. There's an exception, you're allowed to post that content, and that's literally all the content we're posting. So… are we not allowed to post this? Because [Facebook's] rule says you're not allowed to except in this case, which it is key. So what's the deal here? I mean, you can't have every single post be manually vetted by a human, because that's physically impossible. But something's obviously not working.*

Based on his experience with the removal of his top surgery photos, P4 perceived that, despite Facebook's nudity guidelines allowing top surgery photos[22], the platform's algorithmic moderation systems could not actually distinguish between top surgery photos and other kinds of unallowed topless photos. P4 identified both the perceived flaws of Facebook's algorithmic moderation systems and the specific inequities he experienced on Facebook as a transgender man, both of which negatively impacted P4's perception of Facebook's platform spirit and degraded his relationship with the platform.

Other users shared their own negative perceptions of algorithmic identity-related content removals or suppression based on their own negative experiences with platforms' moderation of their content; P8 stated that Instagram's algorithmic removal of their LGBTQIA+ healthcare-related imagery was *"creepy"* and *"invasive,"* and attributed the removals to flaws in Instagram's algorithmic moderation systems' ability to distinguish between images of bodies that either do or do not violate Meta's guidelines on nude and graphic content. When P9 suspected that their trans-related political Facebook posts were being algorithmically suppressed by the platform, they asked their Facebook friends to "like" or otherwise engage with the posts in question; after experiencing a low number of engagements from their Facebook friends, P9 *"vaguely attributed"* the unexpectedly weak Facebook engagement to the presumed *"overaggressive"* algorithmic suppression of their trans-related political posts. Like P4, both P8 and P9 addressed their experiences with identity-related content removals or suppression that they recognized to be incorrect, including the disproportionate moderation they faced specifically as marginalized users posting identity-related content, with P9 taking additional steps to verify whether their identity-related posts were being suppressed. And like P4, P8 and P9's experiences negatively impacted their perceptions of Instagram and Facebook's platform spirit respectively; even if it wasn't the platforms' intent to disproportionately remove marginalized users' content, the fact that participants experienced identity-related removals that they recognized to be incorrect still resulted in them developing negative perceptions of their platforms' spirit. When participants witnessed the incorrect algorithmic suppression of marginalized users' content, they recognized their platforms' algorithmic moderation tools as harming marginalized users and improperly enforcing platform guidelines, resulting in users developing increasingly negative perceptions of their platforms' spirit.

4.1.2 Platform Spirit and Platform Guidelines. Users also generally recognized platform guidelines as either not designed to include marginalized users or not enforced in a way that keeps marginalized users safe and free to express themselves on their platforms, drawn from their personal experiences with inequitable platform policies and enforcement of policies on social media platforms. For example, P6, an Asian nonbinary user who frequently experienced racist abuse, transphobic abuse, and incorrect removals of trans-related content on TikTok and

Facebook, shared their negative perception of Facebook and TikTok based on their negative experiences with the two platforms' inequitable community guidelines: *"I think [poor guidelines] especially are apparent on both TikTok and Facebook. Both apps tolerate white supremacy and protect white supremacists, but don't protect black and brown and intersectional identity creators."* P6 then questioned the purpose of community guidelines that reinforce bigotry and expose marginalized users to harm, asking: *"I understand the need for rules… but when the rules hurt the marginalized but protect literal white supremacists, what are your guidelines actually *doing* other than reinforcing that ideology?"*

Similar to P3 and P4, P6 experienced racist and transphobic abuse from other Facebook and TikTok users, along with incorrect removals of their identity-related Facebook and TikTok content. Like P3 and P4, P6 developed a negative perception of TikTok's and Facebook's platform spirit resulting from their experiences with racist and transphobic abuse on both platforms, as P6's experiences on TikTok and Facebook contradicted their expectation that they may use both platforms as freely and safely as non-marginalized users. After P6 identified TikTok and Facebook's guidelines as enabling white supremacist values, racist abuse, and transphobic abuse, they then recognized the two platforms themselves as tolerant of white supremacy and transphobia, and therefore unsafe for BIPOC users (particularly Black and brown social media users), transgender users, and marginalized users broadly. Negative and painful experiences with guidelines that exclude or outright harm marginalized users lead marginalized users to develop worse perceptions of the platforms as a whole, contributing to their negative perceptions of the platform's spirit.

Several participants also reported not trusting social media guidelines as accurate, helpful guides for behavior on platforms, emphasizing that community guidelines are subject to frequent updates that can be difficult for ordinary users to track. P14 shared that she perceived social media guidelines in general as including *"a lot of sneaky updates and agreements,"* while P15 perceived reading Instagram's guidelines as pointless as *"they're just gonna update them anyways."* P14 and P15 both perceived their platforms' guidelines as updated too frequently and unclearly to accurately guide users' behavior; in response, both participants expressed mistrust of (and a reluctance to read) social media platforms' guidelines as a whole. P14 and P15's mistrust of their platforms' frequently-updated guidelines exacerbated their frustrations with identity-related content removals, contributing to their negative perceptions of their platforms' spirit and degrading their relationships with their platforms.

Overall, participants' perceptions of social media platforms degraded as they experienced or witnessed disproportionate and inequitable removals of identity-related content on those platforms. The participants' experiences with identity-related content removals, along with the moderation-related tools that enable incorrect identity-related removals (such as inaccurate algorithmic moderation tools or unreliable written guidelines), resulted in the participants identifying platforms' moderation of their content as both inequitable and inconsistent. The inequity and inconsistency of both platforms' guidelines and moderation of identity-related content resulted in participants perceiving their platforms as having a negative platform spirit.

Instead of relying on platforms' guidelines that they recognized to be unhelpful or untrustworthy, participants more often used their perceptions of platforms' negative spirit, combined with their personal experiences with identity-related moderation, to develop social media folk theories to guide their behavior on platforms. In what follows, we describe how participants developed folk theories which addressed their negative perceptions of platforms' spirit (including negative perceptions of platforms' algorithmic systems, guidelines, and other affordances), informed by their personal experiences with inequitable identity-related content and account removals.

4.2 Users' Folk Theories

Every participant in our study developed folk theories as sensemaking tools that explain their experiences with and perceptions of identity-related moderation on social media platforms. While some folk theories are unique to

individual users, most folk theories are the complex, layered products of individual experiences with content moderation, personal perceptions of platforms' spirit, and other existing theories within specific identity-based user communities, reflecting similar theorization trends to those presented by DeVito [12, 13]. Nearly all of the participants' folk theories denote a baseline perception within marginalized social media user communities that social media platforms disadvantage marginalized users in some way.

The folk theories participants shared about content removals are explicit, structured theories that draw from negative perceptions of platforms' spirit like those described in section 4.1, where platforms (or their algorithmic moderation systems) are considered at fault for disproportionately removing marginalized users' content. Contrasting past work on platform spirit and folk theories, which found that certain types of folk theories can erode perceptions of platform spirit [11, 12], our study participants drew from their eroded perceptions of platforms' spirit to directly inform their folk theories, *using platform spirit itself as a basis for their subsequent folk theorization.* While some theories refer to marginalized users as a whole, most theories focus on how a specific marginalized user community is impacted by content moderation. In what follows, we describe participants' folk theories in five categories: theories addressing algorithms, theories and perceptions of platform spirit, theories reinforcing other theories, theories addressing unclear platform guidelines, and theories addressing platforms' values.

4.2.1 Theories Addressing Algorithms. Folk theories surrounding content removal and suppression, particularly those associated with algorithmic removals, are often used by marginalized social media users to inform other marginalized users (e.g., in online communities) of how to adjust their behavior to avoid having their own content removed or suppressed on a platform [14]. Even though ordinary users cannot know the exact mechanics of algorithmic content moderation or sorting on a platform, their folk theories show that they do perceive which kinds of activities on a platform are most likely to trigger an algorithmic response. P9, a nonbinary Asian Facebook user, shared several personal folk theories that they adhere to in order to make their content more visible while avoiding incorrect removals:

> *People posting their fundraisers or something say they're finding ways to censor "PayPal" or "CashApp" or whatever. [Many] of my friends say things like "image for attention" or "image for algorithm." I feel like Facebook prioritizes images over other kinds of content... And [other users] will say "please repost this entire post to your own feed instead of just pressing 'share,' because more people will see it that way." Or they'll say, "don't say 'bump' or 'boost' in the comments, because Facebook is trying to suppress those comments. Instead, if you want the post to be seen, write a full sentence or post a GIF in the comments."*

P9 described several folk theories (based on the actions and theories of other Facebook users) about how Facebook's content recommendation algorithms choose content to make more or less visible on the platform (such as prioritizing content that includes images or suppressing posts whose comments include "bump" or "boost"). Though P9 could not know for certain whether each theory about Facebook's algorithms was true, they were confident enough in their folk theories to act on them in order to avoid the suppression of their Facebook content. P9 related their folk theories about algorithmic content suppression to more general folk theories about marginalized user communities being disproportionately impacted by content removal and suppression:

> *Even if it's not intended on the surface to hurt marginalized people, who's going to need to raise money? Or who's feeling impacted by an issue and wants to raise people's awareness? And they have to go through all these hoops in order to make their posts get seen, and even then their post isn't seen the same way that just posting a photo of your pet would be.*

Participants typically formed links between their folk theories and their perceptions of social media platforms' apathy or hostility toward marginalized user groups. P10, a queer man from India, shared his theory that Twitter's algorithmic moderation tools do not properly detect posts that include transphobic slurs in non-English languages,

basing his theory on his past experiences seeing non-English transphobic tweets repeatedly go unremoved from the platform. Like P9, P10 connected his specific theory about non-English transphobic Twitter content to his broader theories about platforms exhibiting apathy toward marginalized user groups, particularly those outside of the US. In this case, his broader theory was that Twitter, a US company, *"does not consider the cultural context of different countries"* while developing its content moderation algorithms, creating a destructive environment for both queer Indian Twitter users and for Twitter users outside the US in general.

Overall, the participants who shared theories about social media algorithms typically saw the theories as evidence that certain groups of users are disadvantaged on the platform in some way, either through disproportionate algorithmic content removal or visibility suppression. Even when a folk theory is meant to inform a marginalized user on how to behave on a platform, it can also reinforce their overall perception that the platform is biased against them to begin with, negatively impacting their perception of the platforms' spirit and degrading their relationship with the platform.

4.2.2 Theories and Perceptions of Platform Spirit. Some participants created folk theories explicitly addressing the perceived reasons for their negative perceptions of a platforms' spirit, acknowledging their negative impression of their social media platforms. The participants who developed these theories typically wanted to understand why a platform would moderate content in a way that violated the participants' expectations for how a fair and equitable platform should moderate content. For example, P12, a Latina Instagram user whose swimwear selfie was removed for allegedly violating community guidelines, felt frustrated when her posts were removed while similar content posted by advertisers remained on the platform:

> *I had one photo taken down [from Instagram]; from what I recall, I think it was seen as "lewd." But I was completely covered up, it was in a bathing suit… They said that it violated community guidelines or something like that. But what I don't understand is that you see ads on Instagram all the time for companies advertising bathing suits or other clothing where the models are scantily clad, and they don't get their posts taken down. So I feel like anything that is considered "lewd" that doesn't make money for a company is seen as "going against community guidelines," so it needs to be removed. That's what it seems like.*

Prior to the removal, P12 had an expectation that her swimwear selfie did not violate Instagram's guidelines and was allowed to be posted on the platform. But not only was P12's swimwear selfie removed, she then noticed that similar swimwear images were not removed from Instagram when posted as corporate or organizational advertisements. P12 found Instagram's removal of her swimwear selfie (while not removing similar images posted by advertisers) to be unfair and inconsistent, contradicting her expectation of how Instagram's community guidelines *should* be enforced. P12 developed two folk theories in response to her selfie removal: a theory that Instagram considers swimwear selfies to be "lewd" content that is subject to removal, and a theory that Instagram allows "lewd" images to be posted as paid advertisements by corporations and organizations, but not by ordinary users. These two theories helped P12 make sense of why her swimwear selfie was removed while similar content posted by paid advertisers is not removed. Theories like P9's and P12's also demonstrated a unique relationship between perceived platform spirit and folk theorization, in that users' perceptions of platform spirit became an information source serving as a basis for their folk theorization, situated alongside and expanding our understanding of the endogenous and exogenous information sources for folk theorization described in earlier work [11]. Like P9 in section 4.2.1, P12's negative perceptions of Instagram's platform spirit deepened while developing her theories, further degrading her relationship with the platform.

4.2.3 Theories Reinforcing Other Theories. Even folk theories by marginalized users that seemingly do not relate to one another, or seem unrelated to marginalization and identity, can intersect with other theories that explicitly relate to those topics. Some participants drew connections between their existing folk theories, even if they

seemed only tangentially related to one another at first, to develop new folk theories addressing aspects of platforms' moderation systems that their previous folk theories did not. Some users then took another step by developing new folk theories based on the connections they perceived between their existing folk theories. For example, P6, a mixed-race nonbinary TikTok user who experienced the removal of a video that they believe did not violate community guidelines, initially theorized that TikTok disproportionately moderates and removes content posted by "small" accounts that do not have a large number of followers on the platform:

> *TikTok will take down videos from small creators where it's kind of a non issue, [it] really doesn't violate anything. But there'll be huge accounts that post, you know, murder scene cleanup videos. And it's like, those are up, that account has a huge following... I think those accounts tend to bring traffic to platforms like TikTok, where they kinda do benefit from all the views. So [platforms] kind of turn a blind eye, whereas with smaller creators it's less consequential.*

P6 then tied this theory to their parallel theory that large TikTok accounts with many followers disproportionately belong to white users, meaning that TikTok's preferential treatment toward "large" accounts translates in practice into preferential treatment and elevated visibility for white TikTok users.

> *[From] what I've seen, TikTok definitely does favor larger white creators. Then they take down a lot of videos of minorities... So many minority creators, Black, Indigenous, and People of Color (BIPOC), get their videos taken down for no reason at all. So it's frustrating to see how this community guideline situation happens. It's definitely kind of shady.*

By combining two folk theories based on observation, other users' experiences, and a general negative perception of TikTok's platform spirit, P6 developed a new folk theory that TikTok's content recommendation algorithm promotes and privileges white creators over BIPOC creators. This new theory validated P6's negative perception of TikTok's platform spirit, helping them explain their perception that TikTok shows preferential treatment to its white content users through its algorithmic content recommendation and moderation systems. Even if P6 did not have an exact understanding of how the algorithms operate, they were confident that their folk theory (based on their observations and existing folk theories about the platform) explains some mechanics of how TikTok's algorithmic content recommendation and moderation systems could disadvantage its BIPOC users.

Like P9 and P12 earlier, P6 drew connections between their existing folk theories related to the disproportionate removal of marginalized users' content (in this case on TikTok); P6 then took another step by developing a new folk theory about TikTok's disproportionate moderation of BIPOC users' content based on the connection they found between their two folk theories. P6's theory reinforced their perception of TikTok being negatively biased against BIPOC content creators like themself, degrading their relationship with the platform.

4.2.4 Theories Addressing Platform Guidelines. Other participants developed folk theories that addressed aspects of platforms' guidelines, such as their lack of clarity, exclusion of marginalized users, and embedded harms toward marginalized users. Participants were particularly likely to theorize about platforms' guidelines if they recognized their platforms' guidelines (and their platforms' subsequent identity-related content removal decisions) as inherently discriminatory against their marginalized identity. P3 gave an example of harm embedded in the wording of social media guidelines while sharing their experience of having a topless selfie removed from Instagram for, in their words, "not having cis male nipples":

> *So many trans creators on Instagram who are banned from the app just for being trans... There is a huge issue with [trans] content being removed. [Instagram's] policy itself is totally biased and skewed towards cisgender and heterosexual people who hold cisgender and heterosexual and white identities - because it was written by them! I don't think that there's necessarily malicious intent, but I do think that there is a consequence to that. It's damaging to people who hold marginalized identities, damaging to their ability to interface with the software [and] with the app, to socialize, and to feel included.*

In this instance, P3's topless selfie was removed, leading them to experience alienation and invalidation on the platform. P3 felt targeted by Instagram's content guidelines and, guided by both their own content removal and witnessing similar removals happen to other nonbinary Instagram users, theorized that Instagram's policies inherently discriminate against non-cis male Instagram users, directly harming trans and nonbinary users as a result. P3's theory would later be explicitly confirmed by the Oversight Board[3] itself, when the Oversight Board overturned Meta's incorrect removal of two trans and nonbinary users' top surgery fundraising posts while officially recommending that Meta clarify its Adult Nudity and Sexual Activity policy to avoid imposing cisnormative views of bodies on transgender and nonbinary users [61]. P3 also theorized that Instagram's guidelines are written by people who hold cisgender, heterosexual, and white identities, and favor cisgender, heterosexual, and white users as a result. Other users shared their own theories about social media platform guidelines based on their content removal experiences; P9 theorized that Facebook's guidelines prohibit criticizing men broadly on the platform, while P8 (a nonbinary healthcare worker) theorized that Instagram's ban on graphic content extends to medical content, limiting the kinds of medical information and resources that can be shared on the platform. Theories regarding platform guidelines were also often based on the participants' difficulty understanding the guidelines themselves; P9 stated that Instagram's community guidelines are *"not user friendly"* and difficult to find on the platform, while P10 expressed frustration with keeping up-to-date on platform guidelines that are updated often but do not clearly communicate its changes. Participants responded to their uncertainties and harms they experienced due to platforms' guidelines by relying on their folk theories about guidelines to guide their behavior on platforms instead of the guidelines themselves, (a continuation of the dynamic discussed in section 4.1.2). In turn, the perceived need to theorize about guidelines instead of trusting them as written, and doubt that the guidelines would equally include marginalized users, degraded participants' perceptions of their platforms' spirit and their relationships with their platforms.

4.2.5 Theories Addressing Platforms' Values. Some participants shared folk theories addressing the perceived relationship between social media platforms' public stances on social issues related to marginalized communities and platforms' disproportionate moderation of marginalized users in practice. In general, participants' theories about platforms' values indicated their desire to understand whether platforms' public "support" of their communities was sincere or performative, and (by extension) whether they could trust "supportive" platforms to treat their marginalized users equitably. Some of these theories addressed platforms' values related to marginalized users broadly; P9 stated that Facebook and Instagram *"don't actually care"* about the moderation struggles faced by marginalized users, while P11 shared their perception that Twitter is *"apathetic"* toward the harassment faced by marginalized users on their platform. Other participants shared theories addressing the motivations behind a platform acknowledging a social issue publicly. For example, participants noted that, while some platforms enabled users to change their profile pictures to a rainbow theme for Pride Month, or publicly acknowledged the Black Lives Matter (BLM) movement, these actions conflicted with the same platforms' disproportionate removals of marginalized users' content. As a result of this conflict, the majority of participants expressed distrust of platforms that engage with social issues in this way (P3 called the phenomenon a *"very placating and performative gesture"*), and developed folk theories associating platforms' engagement with social issues with a desire to appear socially conscious to the general public instead of a sincere intent to treat their marginalized users in an equitable way.

As platforms' public social stances can conflict with their actual moderation practices, marginalized social media users often face unequal treatment on platforms that publicly claim to support them. For example, P1, a

[3]The Oversight Board is an independent governing body that oversees Meta's content moderation decisions and appeal cases; the Oversight Board can make policy recommendations for Meta and overturn incorrect content moderation decisions that violate Meta's Community Standards [60].

Black LinkedIn user whose content about BLM was removed from the platform, shared her theories about the relationship between platforms and social issues:

> *I noticed that with LinkedIn... at least on my feed, they have been doing a lot of changes that completely remove posts and [accounts] calling out white supremacy. Like, comments about BLM will get automatically deleted... If you're deleting people calling out injustice and asking people to be held accountable, you're a hypocrite. You're a hypocritical company.*

P1 witnessed LinkedIn introduce BLM-themed graphical assets (such as profile banners) to their website for users to freely use; because of this, she assumed that she would be welcome to discuss the BLM movement on LinkedIn as well. However, she then witnessed LinkedIn algorithmically removing content that explicitly mentioned BLM (including her own posts), despite LinkedIn's alleged support of that very same movement. As a result, P1 developed a folk theory that tied these contradictory observations together, theorizing that "LinkedIn is willing to perform superficial acknowledgement of the BLM movement, but is unwilling to host visible discussions about the topic on their platform." P1's folk theory tied her experience on LinkedIn to her overall understanding of how antiblackness operates in the corporate world, even behind the smokescreen of performative allyship. Participants like P1 ultimately theorized that platforms' engagement with social issues is typically performative, and that a platforms' publicly stated "support" of social issues and marginalized groups does not translate in practice into equitable treatment of their marginalized users. These theories reinforced the participants' negative perceptions of their platforms' spirit and degraded their relationships with their platforms. For P1, her experiences and theories caused her to develop a deeply negative perception of LinkedIn's platform spirit and to consider leaving LinkedIn for other professional networking platforms.

Overall, participants shared a variety of folk theories about social media platforms and their moderation practices. These theories ranged in topic; some addressed platforms' mechanics (such as their algorithmic moderation and recommendation systems), while others addressed platforms' guidelines and the sincerity of platforms' publicly stated social values. Some participants also theorized about the reasons for their negative perceptions of a platforms' spirit, or used their existing folk theories to develop new theories about their platforms. Participants typically theorized about the elements of their social media platforms' moderation systems that drove their negative perception of the platforms' spirit; said theories could reinforce (or even magnify) participants' negative perceptions of platforms' spirit, accelerating their degrading relationships with their platforms.

In what follows, we describe how participants responded to their folk theories about social media platforms that they perceive to disproportionately moderate marginalized users' content, such as changing their behavior on platforms, reducing their use of platforms, or leaving platforms entirely.

4.3 User Behaviors/Responses to Theories

After developing folk theories based on their negative perceptions of platforms' spirit and their own personal experiences with content moderation, participants typically adjusted their behaviors and decision-making on platforms in response to those theories. All participants perceived marginalized identity-related social media content as likely to face incorrect suppression and removal regardless of the platform they were posted on or the platforms' guidelines themselves. Notably, participants' perceptions of platform guidelines being unreliable for marginalized users encouraged them to rely on their folk theories to guide their behavior and decision-making on platforms instead of relying on the platforms' guidelines themselves.

4.3.1 Using Coded Language to Avoid Incorrect Algorithmic Moderation. Substituting coded language for explicitly identity-related vocabulary was the most common behavioral response to folk theories about platforms reported by participants. For example, participants who experienced the algorithmic removal of posts including identity-related terminology (like P1 in section 4.2.5, whose LinkedIn posts including the term "BLM" were removed from the platform) theorized that certain identity-related words and phrases flag platforms' algorithmic moderation tools

and result in those posts (and possibly the users' accounts) being incorrectly removed. In response, participants described that they generally avoided including identity-related phrases and words on their social media posts; instead, they substituted slang, deliberate misspellings, and abbreviations to encrypt the meaning of more explicitly identity-specific words that they theorized would attract algorithmic removals. For example, P7 stated that they misspelled the names of political figures that they criticized on Twitter to avoid being flagged for "harassment," while P9 reported avoiding certain phrases while discussing trans-related issues on Facebook to avoid having their posts falsely removed for "hate speech." Several users also reported limiting the kinds of images they posted due to similar theories involving algorithmic image moderation, such as P4 avoiding posting gender affirming surgery images on Facebook even though those images are explicitly permitted by Facebook's community guidelines.

Participants' perceived need to avoid using identity-specific terms that may trigger algorithmic removals reflects similar findings from past studies related to marginalized users' theories about platforms suppressing identity-specific speech [46]; the perceived need to avoid using identity-related words and phrases not only influenced participants' use of language on platforms, but also reinforced their negative perceptions of their social media platforms' spirit. P7 stated that their perceived need to obscure the names of political figures on Twitter made them feel *"unsafe"* and *"threatened"* on the platform, while P18 stated that he felt *"rattled"* and *"censored"* by the algorithmic keyword flagging that he theorized took place with his Twitter and Facebook posts. P9 stated that the need to substitute certain "flagged" terms could create barriers of communication with other marginalized users, stating that *"some people could understand what I'm talking about if I use these euphemisms, but others may be confused."* Ultimately, marginalized users who theorize that platforms suppress identity-related terms are likely to dodge that censorship in ways that the platform may not have intended. The perceived need to substitute coded language for explicitly identity-related phrases can also negatively impact marginalized users' perceptions of platforms' spirit, degrading both their relationship with their platforms and their willingness to continue using their platforms in the long-term.

4.3.2 Leaving or Reducing Use of the Platform. Many participants reported leaving or significantly reducing their use of platforms after developing theories addressing how and why their identity-related content removals took place. For example, P17 is a Black Instagram user whose post about the Black Lives Matter movement received backlash and harassing comments from several anti-BLM users before being removed by Instagram itself. P17 felt frustrated when Instagram's removal alert did not state which rule her post allegedly violated; she speculated that her post may not have violated Instagram's guidelines at all, and instead wondered if anti-BLM users mass-reported her post to have it algorithmically removed instead. P17 eventually theorized that "Instagram algorithmically removes posts that do not violate community guidelines so long as they receive a certain number of reports – allowing the report feature to be abused for bigotry and harassment." This theory, along with the removal itself, left P17 with a deeply negative perception of Instagram's platform spirit and an unwillingness to keep using the platform:

> *I haven't really posted since the removal, even though there have been other issues to discuss [involving the BLM movement]. I think it deterred me from posting, not because I don't want to get the message out there, but because… is this really important to Instagram? Are their values aligned with my values? Are they going to delete my post again? So yeah, I just… haven't posted since then.*

Here, P17 reveals her negative perception of Instagram's platform spirit, informed by her folk theory and frustrating experiences surrounding the removal of her BLM-related post. She also revealed that she no longer posts on Instagram as a result of this incident, as she no longer perceives Instagram's values as aligning with her own. Other participants also reported leaving platforms after experiencing identity-related content removals; after experiencing persistent harassing comments and false reports of her Facebook selfies, P13 theorized that Facebook does not take a strong stance against cyberbullying and harassment, emphasizing other users' persistent

abuse of the report feature as evidence for her theory. She acted on her theory by deactivating her Facebook account, stating that she's *"done with Facebook"* and has no intention of returning to the platform. P24 also stated that he no longer posted on Reddit after his post about experiencing ADHD was removed, while P16 no longer posted on Instagram after incorrectly having her selfie removed for "nudity."

Other participants reported significantly reducing their use of platforms or avoiding specific features: P6 significantly reduced the number of trans-related TikToks they posted to avoid having their videos removed from the platform again, while P15 began only posting on Instagram's "Stories" feature after theorizing that her non-Stories posts would continue to be algorithmically removed. P1, whose experiences on LinkedIn were discussed in Section 4.2.5, began exploring alternative professional networking platforms where she could continue her online networking without experiencing the content removals that she did on LinkedIn. In the same way that folk theories guide the behavior of users who remain on a platform, folk theories can also guide users into reducing their use of a platform or leaving it entirely, having decided to step away from platforms that they perceive as targeting, disproportionately moderating, and harming users like themselves.

Overall, participants who theorized that social media platforms disproportionately remove marginalized users' identity-related content (and developed negative perceptions of those platforms' spirit as a result) responded to their theories in a variety of ways. Some participants chose to substitute coded language for identity-related speech on their platforms despite identity-related speech being allowed on those platforms, expressing fear that openly using identity-related speech would result in even more of their content being algorithmically removed. Other participants acted on their theories by reducing their use of their platforms or leaving them entirely, having determined based on their theories that their social media platforms do not provide safe, welcoming, and equitable experiences for marginalized social media users.

5 DISCUSSION

We have explored and described how marginalized social media users draw from their perceptions of platform spirit to develop folk theories about social media platforms and their content moderation practices. **RQ1** asked how marginalized users' perceptions of platforms' spirit influence their folk theorization about social media platforms and their content moderation practices. We found throughout section 4.1 that our participants developed negative perceptions of platforms' spirit as they experienced or witnessed identity-related content removals that they perceived to be incorrect. Participants held particularly negative perceptions of platforms' algorithmic moderation and content sorting systems (perceived to be inconsistent and biased against marginalized users' content) and written guidelines (perceived to exclude or outright harm marginalized users in their wording), both of which worsened participants' perceptions of their platforms' spirit.

We then explored in sections 4.2 how participants created theories addressing the aspects of social media platforms that fueled their negative perceptions of their platforms' spirit. Participants' theories addressed a broad range of topics related to social media platforms and content moderation, such as theories about platforms' algorithmic moderation systems, platforms' guidelines, platforms' values, or related to the participants' perception of a platform's spirit; some participants also shared folk theories related to their other existing theories about platforms, and even synthesized multiple theories in order to develop new ones. **RQ2** asked how marginalized social media users adjust their online behavior in response to their folk theories. We found throughout section 4.3 that participants who theorized about platforms' algorithmic moderation systems often responded by substituting coded, indirect language for explicitly identity-related speech on platforms to avoid accidentally having more of their content algorithmically removed. Many participants also described reducing their use of platforms in response to their theories – or leaving their platforms entirely.

Drawing on our results, we next discuss how marginalized users' content moderation folk theories can reveal platform design issues that enable the disproportionate moderation and harm they experience on platforms.

We extend prior literature in three ways: expanding on folk theorization in the content moderation context, expanding on the relationship between folk theorization and platform spirit while demonstrating bidirectional interplay between the two concepts that differs from the unidirectional relationships explored in past work, and exploring ways in which marginalized users' folk theories can be used to identify ways in which platforms' tools and affordances are not working as intended. Past literature has explored the use of folk theorization as a sensemaking tool among both general social media users [17, 21, 26] and specific marginalized groups such as online LGBTQ+ content creators [11, 12]. We expand on this work by examining these phenomena in a content moderation context; we explore how marginalized social media users' negative perceptions of platforms can lead them to develop folk theories to navigate the threats they encounter on those platforms, and how those theories pinpoint the weaknesses in platforms' design that enabled the mistreatment of marginalized users to begin with.

Past literature has focused on the use of folk theories to make sense of algorithmic moderation, including social media users' ability to create complex theories about algorithmic systems [17] that are particularly opaque and otherwise difficult for users to understand [11, 46]. Our research extends this work by providing further insight on marginalized social media users' perceptions of algorithmic moderation systems, including how their negative perceptions of platforms' spirit inform their folk theories, how they use their folk theories to explain algorithmic content removals, how they adjust their behavior in response to algorithmic moderation, and how they interpret algorithmic moderation tools as inherently biased against them.

Past literature has highlighted how the "affective dimensions" of content moderation [57] inform users' desire to theorize about content moderation systems and to act on their theories through changing their behavior on platforms, participating in platform policy design, or even through acts of civil disobedience [46, 57]. We extend this past work by specifically exploring how marginalized users' negative perceptions of platforms' spirit, guided by their experiences with content moderation, informs marginalized users' folk theories about, decision-making on, and degrading relationships with said platforms. In contrast to past work, participants in our study did not emphasize a desire to act on their folk theories by involving themselves in platform governance or civil disobedience; instead, participants specifically used their theories to either guide their use of identity-specific language on platforms or to decide whether to reduce their use of platforms or leave them entirely.

Past work by DeVito focused on how negative or "demotivational" folk theories erode users' perceptions of platform spirit [12], and how negative perceptions of platform spirit lower users' willingness to adapt their behavior in order to remain on platforms [11]. Our findings add a new dimension to the cycle demonstrated in DeVito's past work – while DeVito demonstrates transfeminine TikTok users' negative (or "demotivational") folk theories about the platform degrading their perception of TikTok's platform spirit [12], our study additionally shows how marginalized social media users' negative perceptions of platforms' spirit (specifically guided by their disproportionate content moderation experiences) directly informs their negative folk theories about said platforms, demonstrating that the relationship between platform spirit and folk theorization can manifest in both directions.

In what follows, we first discuss how folk theories that cannot be entirely proven true are still valuable tools that accurately measure social media users' sentiments and perceptions of their platforms. We then make recommendations for how social media platforms can draw from marginalized users' folk theories as troubleshooting tools that can help quickly identify and repair elements of platform design, including moderation tools, that are not behaving as intended and are inadvertently harming marginalized users and driving them away from the platform.

5.1 Do Folk Theories Need to be True to be Valuable?

We argue that social media folk theories that cannot be entirely proven are still particularly valuable to marginalized users who are disproportionately likely to face structural obstacles while using these services. Marginalized

social media users do not need a strict system-level knowledge of how a platform's moderation systems work, or even a complete knowledge of a platform's community guidelines, to recognize when content moderation systems disproportionately harm them or their community. Marginalized users' folk theories about moderation systems are not fictitious, but are instead products of empirical evidence informed by both their specific lived experiences with discriminatory content moderation systems and by their personal insight drawn from pervasive offline experiences with systemic oppression and exclusion. These lived experiences inform folk theories that meaningfully address marginalized users' negative and discriminatory experiences with moderation systems where direct insight into those systems is not possible, while situating their experiences in the context of their overall experiences with marginalization. In turn, the theories that marginalized users develop to make sense of their experiences, to help one another continue to participate on these platforms, and even to inform direct action taken against platforms to draw attention to their inequitable content moderation practices, do not require a strict mechanical knowledge of how platforms actually moderate. Instead, folk theory formation only requires that marginalized users have something about a platform that they need explained. Negative perceptions of the platform's spirit are a strong motivator for this need, as they are generally driven by marginalized users' own negative identity-related treatment on the platform, and may target a specific platform feature, moderation practice, or even the platform as a whole.

Marginalized users' folk theories pinpoint the specific elements of platforms' design that expose them to harm; this allows their folk theories to be useful tools that reveal the specific elements of a platform's design that do not work as the user would expect, and thus inadvertently expose marginalized users to harm. This reinforces that it does not really matter if marginalized social media users know how the site's algorithms really work, or the exact system-level code that results in platforms suppressing their content or exposing them to harm. So long as marginalized users' recognize that platforms disproportionately censor or harm them in some way, they will begin developing theories that explain why their experiences on platforms have been so bad and guide their future behavior on those platforms (assuming they stay on those platforms at all). The existence of common folk theories addressing struggles within specific marginalized user communities indicates that platforms likely disproportionately harm those communities in some way.

Many folk theories exist because marginalized communities do not trust social media platforms to treat them equitably and to keep them safe from abuse. Our interview study reaffirmed that this mistrust is well-founded, as marginalized social media users frequently encounter abuse and unfair moderation on social media that informs their mistrust of platforms (and their need to theorize to begin with). The existence of these negatively-tinged folk theories themselves are proof enough that social media platforms fail to create an environment where marginalized users feel they can participate equally and safely. The folk theories' specifics, accuracy, and provability are less important than the fact that they exist in the first place.

5.2 How Can Platforms Respond to Folk Theories?

We argued above that marginalized users' folk theories, even when developed without a strict and accurate algorithmic understanding of platforms' moderation systems, still effectively identify elements of platform design that do not work as intended and disproportionately harm marginalized users. These theories are perceived by users to be reliable; all study participants used their folk theories to guide their behaviors on those platforms, including those whose folk theories guided them into leaving the platform entirely. Platforms can gain valuable insight by analyzing marginalized users' folk theories to assess their overall sentiments of the platform, as well as to assess whether marginalized users are particularly negatively impacted by a specific platform tool or affordance. Some social media platforms have already explored strategies for developing more inclusive platform policies and moderation practices; for example, Grindr's whitepaper on gender-inclusive moderation provides guidance on incorporating gender-inclusive platform features (such as inclusive gender and pronoun options) and

moderation practices (such as incorporating "gender-free" image moderation policies where possible) to develop more equitable, inclusive user experiences for a gender-diverse user community [39]. Platforms like Grindr who wish to improve the inclusivity of their policies and affordances could particularly benefit from analyzing their marginalized users' folk theories, allowing them to more efficiently identify the ways in which marginalized users are currently alienated or harmed on their platforms. Below, we explore several ways in which platforms can utilize marginalized users' folk theories as troubleshooting and feature-testing tools, allowing them to more efficiently repair the design problems that caused marginalized users to develop these theories to begin with.

5.2.1 Folk Theories as Troubleshooting Tools. We argue that folk theories can be utilized by social media platforms as troubleshooting tools that allow them to quickly assess users' negative perceptions of their platform and its affordances, as well as where those negative perceptions are coming from. This approach can be particularly valuable for platforms that may experience concern regarding user retention. We found in our study that marginalized users who left social media platforms developed theories that addressed the perceived causes of their frustrations on the platform; users who left often did so in response to their folk theories, which in turn existed because they faced negative experiences on the platform to begin with. P13's experience on Facebook (described in section 4.3.2) demonstrates this process: as described in the results, P13 chose to leave Facebook in response to her theory that Facebook does not take a strong stance against harassment, a theory that she created in response to her experiences with persistent harassment and false reports from abusive users. A platform such as Facebook could utilize theories like P13's to quickly investigate and identify ways in which their platform and its tools are not working as intended; in a case like P13's, platform responses could include identifying the report feature's flaws and investigating ways to prevent it from being used for harassment by abusive users. The platform could even use P13's experience specifically to investigate what went wrong, how, and why.

Marginalized users' folk theories, even compared to non-marginalized users' theories, could be particularly useful for identifying flaws in moderation tools and systems, as marginalized users are uniquely likely to be impacted by flawed moderation tools and systems, resulting in their disproportionate rates of moderation and content removals [37]. If a platform is facing negative perceptions from and degrading relationships with their users, assessing marginalized users' folk theories may be an efficient way to pinpoint the causes of users' negative sentiment, allowing platforms to more easily identify and solve its own problems while preventing the loss of future users.

5.2.2 Folk Theories as Proactive Feature-Testing Tools. Similar to how folk theories can be used to troubleshoot immediate problems with a platforms' design, they can also be integrated into the platform's general toolkit and can be regularly utilized to assess users' changing sentiments toward the platform and its features. Though folk theories are often discussed in a reactive context, this approach could allow folk theories to act as proactive feature-testing tools that can measure users' sentiments toward new platform tools and affordances before they are exposed to the entire user community. For example, platforms could integrate their assessment of users' folk theories into A/B testing, inviting users to share their folk theories during A/B testing and measuring how marginalized users perceive new features while watching for negative theories that indicate that the features are unpopular, not functioning as intended, or possibly harming users. Folk theories can also be used outside the feature-testing context as tools to periodically test the platform's current users' sentiments and to see whether any existing aspect of the platform is being perceived negatively. As described throughout the results, the participants in our study frequently perceived that social media platforms did not take marginalized users' frustrations seriously, with users like P9 and P11 theorizing that the platforms are apathetic toward the struggles and harm experienced by marginalized users. As a result, we argue that it is particularly important to integrate marginalized users' folk theories into platforms' regular testing of features and affordances. Integrating these folk theories could allow platforms to ensure that their new tools and affordances do not harm or disproportionately trouble their marginalized users before their official launch. Utilizing marginalized users' folk theories while generally

testing user sentiments on the platform can also reveal whether specific marginalized user communities (or marginalized users broadly) feel negatively about the platform relative to their non-marginalized peers, and which exact elements of the platform lead their frustration. Both of these angles can allow platforms to quickly identify and resolve platform design issues that disproportionately impact marginalized users, preventing further harm while ensuring that the platform's tools work as intended for all users.

5.3 Limitations

We acknowledge the demographic limitations encountered during this interview study. We achieved our goal to speak with marginalized social media users broadly; however, certain user demographics were underrepresented in our participant group. Of the twenty four interview participants, only six of the interviewees were men, including only one Black man. The research team kept in mind that women participants disproportionately reported having their selfies algorithmically removed based on sexual content and nudity guidelines; it is possible that the lack of male participants reflects content moderation trends that give "cis-passing" male users more freedom to post images of their own bodies than other users. Future research could seek more insight into how content moderation and removal policies specifically affect men from different demographics and what the treatment of marginalized men on social media looks like in practice.

We also acknowledge that the research study primarily reflects the experience of social media users from the United States. Twenty one interviewees were located in the United States, and the other three interviewees were located in India. The interviewees from India offered unique insight that reflects a very different set of government policies surrounding content moderation practices and online speech compared to the United States. Future research could focus on what content moderation policies and online marginalization looks like in other specific countries, and whether users in those countries also utilize folk theories to help them navigate different online realities.

6 CONCLUSION

We contributed an overview of how marginalized social media users create and use folk theories as sense-making and behavioral guidance tools to navigate the disproportionate content and account removals that they experience on social media platforms. We explored how marginalized users create folk theories to address their negative perceptions of platforms' spirit after their experiences with content removal and account bans, and how these theories can reinforce their existing negative perceptions of platforms' spirit. We also examined how marginalized users use their theories to guide their decision-making on platforms, including whether they decide to continue using platforms at all, drawing on folk theories derived from their observations and interactions with content moderation rather than trusting that the guidelines and moderation on a platform will treat them fairly. We then explore the potential use of marginalized users' folk theories as troubleshooting tools that can quickly pinpoint aspects of platform design that may not be functioning as intended, resulting in user dissatisfaction. We also explored ways that platforms can potentially use marginalized users' folk theories as proactive feature testing tools to measure users' sentiments regarding new platform features and affordances before deploying them to a wider audience. We argue that using marginalized users' folk theories as troubleshooting and feature-testing tools could allow social media platforms to prevent unnecessary user dissatisfaction, as well as preventing the potential loss of users from their platforms. We hope this paper acknowledges and validates the experiences of marginalized social media users whose content has been unequally suppressed and removed. We also hope that social media platforms learn from these prevailing theories and the conditions that lead to their creation, and pursue changes in community guideline design and content moderation practices that lead to safer, kinder, and more equitable social media experiences for marginalized people.

ACKNOWLEDGMENTS

We thank our study participants for sharing their insights and experiences with us. We thank the members of the Community Research on Identity and Technology (CRIT) Lab at the University of Michigan School of Information (UMSI) for their helpful feedback and comments on this work. We also thank our anonymous reviewers for their constructive comments that improved this work. This work was supported by the National Science Foundation grant #1942125.

REFERENCES

[1] Julia Alexander. 2019. LGBTQ YouTubers are suing YouTube over alleged discrimination. https://www.theverge.com/2019/8/14/20805283/lgbtq-youtuber-lawsuit-discrimination-alleged-video-recommendations-demonetization

[2] 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

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

[4] Elena Botella. 2019. TikTok Admits It Suppressed Videos by Disabled, Queer, and Fat Creators. https://slate.com/technology/2019/12/tiktok-disabled-users-videos-suppressed.html

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[6] Erik Calleberg. [n. d.]. *Making Content Moderation Less Frustrating: How Do Users Experience Explanatory Human and AI Moderation Messages.* Ph. D. Dissertation. Södertörn University, School of Natural Sciences, Technology and Environmental Studies, Media Technology. http://sh.diva-portal.org/smash/record.jsf?pid=diva2%3A1576614&dswid=5032

[7] Christine L. Cook, Aashka Patel, and Donghee Yvette Wohn. 2021. Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms. *Frontiers in Human Dynamics* 3 (Feb. 2021), 626409. https://doi.org/10.3389/fhumd.2021.626409

[8] Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory.* SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States. https://doi.org/10.4135/9781452230153

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

[10] Amanda L. L. Cullen and Bonnie Ruberg. 2019. Necklines and 'naughty bits': constructing and regulating bodies in live streaming community guidelines. In *Proceedings of the 14th International Conference on the Foundations of Digital Games.* ACM, San Luis Obispo California USA, 1–8. https://doi.org/10.1145/3337722.3337754

[11] Michael Ann DeVito. 2021. Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–38. https://doi.org/10.1145/3476080

[12] Michael Ann DeVito. 2022. How Transfeminine TikTok Creators Navigate the Algorithmic Trap of Visibility Via Folk Theorization. (2022).

[13] Michael Ann DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and What it Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173694

[14] Michael Ann DeVito, Jeffrey T. Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 2018. The Algorithm and the User: How Can HCI Use Lay Understandings of Algorithmic Systems?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, Montreal QC Canada, 1–6. https://doi.org/10.1145/3170427.3186320

[15] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25, 2 (April 2021), 700–732. https://doi.org/10.1007/s12119-020-09790-w

[16] 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.

[17] Leyla Dogruel. 2021. Folk theories of algorithmic operations during Internet use: A mixed methods study. *The Information Society* 37, 5 (Oct. 2021), 287–298. https://doi.org/10.1080/01972243.2021.1949768

[18] Nick Drewe. 2016. The Hilarious List Of Hashtags Instagram Won't Let You Search. http://thedatapack.com/banned-instagram-hashtags-update/

[19] Emily Dreyfuss. 2018. Twitter Is Indeed Toxic for Women, Amnesty Report Says. https://www.wired.com/story/amnesty-report-twitter-abuse-women/

[20] Brooke Erin Duffy and Colten Meisner. 2022. Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society* (July 2022), 016344372211119. https://doi.org/10.1177/01634437221111923

[21] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 2371–2382. https://doi.org/10.1145/2858036.2858494

[22] Facebook. 2021. Adult Nudity and Sexual Activity. https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity/?from=https%3A%2F%2Fm.facebook.com%2Fcommunitystandards%2Fadult_nudity_sexual_activity%2F%3Fprivacy_mutation_token%3DeyJ0eXBlIjowLCJjcmVhdGlvbl90aW1lIjoxNjM3MTU2Mjc3LCJjYWxsc2l0ZV9pZCI6MTUwODA5MTU3OTM3NDQ1OX0%253D%26__m_async_page__%26__big_pipe_on__%26fb_dtsg_ag%3DAQwCeH0YZhbwj16xn88Fks9UHTnrTkr9xlge52JYlpiJuYGu%253A34%253A1624034963%26jazoest%3D25008

[23] Facebook. 2022. Hate Speech. https://transparency.fb.com/policies/community-standards/hate-speech/

[24] 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), 1–28. https://doi.org/10.1145/3392845

[25] 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

[26] Megan French and Jeff Hancock. 2017. What's the Folk Theory? Reasoning About Cyber-Social Systems. *SSRN Electronic Journal* (2017). https://doi.org/10.2139/ssrn.2910571

[27] Jason Gallo and Clare Cho. 2021. *Social Media: Misinformation and Content Moderation Issues for Congress.* Technical Report R46662. Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R46662

[28] Susan A. Gelman and Cristine H. Legare. 2011. Concepts and Folk Theories. *Annual Review of Anthropology* 40, 1 (Oct. 2011), 379–398. https://doi.org/10.1146/annurev-anthro-081309-145822

[29] Dedre Gentner. 2001. *Mental Models, Psychology of.* 9683–9687. https://doi.org/10.1016/B0-08-043076-7/01487-X

[30] Ysabel Gerrard. 2020. Social media content moderation: six opportunities for feminist intervention. *Feminist Media Studies* 20, 5 (July 2020), 748–751. https://doi.org/10.1080/14680777.2020.1783807

[31] Shirin Ghaffary. 2021. How TikTok's hate speech detection tool set off a debate about racial bias on the app. https://www.vox.com/recode/2021/7/7/22566017/tiktok-black-creators-ziggi-tyler-debate-about-black-lives-matter-racial-bias-social-media

[32] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (illustrated edition ed.). Yale University Press, New Haven.

[33] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (July 2020), 205395172094323. https://doi.org/10.1177/2053951720943234

[34] GLAAD. 2021. GLAAD's Social Media Safety Index. https://www.glaad.org/blog/glaads-social-media-safety-index

[35] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (Jan. 2020), 205395171989794. https://doi.org/10.1177/2053951719897945

[36] 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

[37] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–35. https://doi.org/10.1145/3479610

[38] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv:1702.08138 [cs]* (Feb. 2017). http://arxiv.org/abs/1702.08138 arXiv: 1702.08138.

[39] Alice Hunsberger, Vanity Brown, and Lily Galib. 2021. Best Practices for Gender-Inclusive Content Moderation. https://static1.squarespace.com/static/5f7cf4654534a21e8041006b/t/61773f8407378d02feeb6f0e/1635205002056/CX-White+Paper-2021.pdf

[40] Irmi Jenzen, Olu; Karl. 2014. Make, Share, Care: Social Media and LGBTQ Youth Engagement. (2014). https://doi.org/10.7264/N39P2ZX3 Publisher: University of Oregon Libraries.

[41] Olu Jenzen. 2017. Trans youth and social media: moving between counterpublics and the wider web. *Gender, Place & Culture* 24, 11 (Nov. 2017), 1626–1641. https://doi.org/10.1080/0966369X.2017.1396204

[42] 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), 1–33. https://doi.org/10.1145/3359294

[43] Jialun 'Aaron' Jiang, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2020. Characterizing Community Guidelines on Social Media Platforms. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. ACM, Virtual Event USA, 287–291. https://doi.org/10.1145/3406865.3418312

[44] 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

[45] 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
[46] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–44. https://doi.org/10.1145/3476046
[47] Knight Foundation. 2020. Americans Support Free Speech Online but Want More Action to Curb Harmful Content. https://knightfoundation.org/press/releases/americans-support-free-speech-online-but-want-more-action-to-curb-harmful-content/
[48] Sam Levin. 2016. Facebook temporarily blocks Black Lives Matter activist after he posts racist email. https://www.theguardian.com/technology/2016/sep/12/facebook-blocks-shaun-king-black-lives-matter
[49] Leanna Lucero. 2017. Safe spaces in online places: social media and LGBTQ youth. *Multicultural Education Review* 9, 2 (April 2017), 117–128. https://doi.org/10.1080/2005615X.2017.1313482
[50] João Carlos Magalhães and Christian Katzenbach. 2020. Coronavirus and the frailness of platform governance. *Internet Policy Review* 9 (March 2020). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-68143-2
[51] Aida E Manduley, Andrea Mertens, Iradele Plante, and Anjum Sultana. 2018. The role of social media in sex education: Dispatches from queer, trans, and racialized communities. *Feminism & Psychology* 28, 1 (Feb. 2018), 152–170. https://doi.org/10.1177/0959353517717751
[52] Brandeis Marshall. 2021. *Algorithmic misogynoir in content moderation practice.* Technical Report. Heinrich-Böll-Stiftung. 17 pages.
[53] J. Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (May 2019), 9785–9789. https://doi.org/10.1073/pnas.1813486116
[54] Bharat Mehra, Cecelia Merkel, and Ann Peterson Bishop. 2004. The internet for empowerment of minority and marginalized users. *New Media & Society* 6, 6 (Dec. 2004), 781–802. https://doi.org/10.1177/146144804047513
[55] Ryan A. Miller. 2017. "My Voice Is Definitely Strongest in Online Communities": Students Using Social Media for Queer and Disability Identity-Making. *Journal of College Student Development* 58, 4 (2017), 509–525. https://doi.org/10.1353/csd.2017.0040
[56] Adam Mosseri. 2020. An Update on Our Equity Work. https://about.instagram.com/blog/announcements/updates-on-our-equity-work
[57] 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
[58] Jibon Naher, An Taehyeon, and Kim Juho. 2019. Improving Users' Algorithmic Understandability and Trust in Content Moderation. Association for Computing Machinery. https://kixlab.github.io/website-files/2019/cscw2019-workshop-ContestabilityDesign-paper.pdf
[59] Fayika Farhat Nova, Pratyasha Saha, and Shion Guha. 2022. Understanding Online Harassment and Safety Concerns of Marginalized LGBTQ+ Populations on Social Media in Bangladesh. In *International Conference on Information & Communication Technologies and Development 2022*. ACM, Seattle WA USA, 1–5. https://doi.org/10.1145/3572334.3572396
[60] Oversight Board. 2022. Appeal to shape the future of Facebook and Instagram. https://www.oversightboard.com/appeals-process/
[61] Oversight Board. 2023. Oversight Board overturns Meta's original decisions in the "Gender identity and nudity" cases. https://www.oversightboard.com/news/1214820616135890-oversight-board-overturns-meta-s-original-decisions-in-the-gender-identity-and-nudity-cases/
[62] Sarah T. Roberts. 2018. Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday* (March 2018). https://doi.org/10.5210/fm.v23i3.8283
[63] Adi Robertson. 2019. TikTok prevented disabled users' videos from showing up in feeds. https://www.theverge.com/2019/12/2/20991843/tiktok-bytedance-platform-disabled-autism-lgbt-fat-user-algorithm-reach-limit
[64] Nick Seaver, Janet Vertesi, and David Ribes. 2019. Knowing Algorithms. In *DigitalSTS: A Field Guide for Science & Technology Studies*. Princeton University Press, 412–422. https://digitalsts.net/wp-content/uploads/2019/11/26_digitalSTS_Knowing-Algorithms.pdf
[65] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 111–125. https://doi.org/10.1145/2998181.2998277
[66] Shakira Smith, Oliver L Haimson, Claire Fitzsimmons, and Nikki Echarte Brown. 2021. Censorship of Marginalized Communities on Instagram, 2021. *Salty* (Sept. 2021). https://saltyworld.net/product/exclusive-report-censorship-of-marginalized-communities-on-instagram-2021-pdf-download/
[67] Robin Stevens, Stacia Gilliard-Matthews, Jamie Dunaev, Marcus K Woods, and Bridgette M Brawner. 2017. The digital hood: Social media use among youth in disadvantaged neighborhoods. *New Media & Society* 19, 6 (June 2017), 950–967. https://doi.org/10.1177/1461444815625941
[68] Nicolas P. Suzor. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge University Press. Google-Books-ID: EjGdDwAAQBAJ.
[69] 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.
[70] TikTok. 2022. Community Guidelines. https://www.tiktok.com/community-guidelines?lang=en
[71] Twitter. 2022. Hateful conduct policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[72] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. https://doi.org/10.1145/3415238

[73] Jillian York and Karen Gullo. 2018. Offline/Online Project Highlights How the Oppression Marginalized Communities Face in the Real World Follows Them Online. https://www.eff.org/deeplinks/2018/03/offlineonline-project-highlights-how-oppression-marginalized-communities-face-real

[74] Rachel Young, Volha Kananovich, and Brett G. Johnson. 2021. Young Adults' Folk Theories of How Social Media Harms Its Users. *Mass Communication and Society* (Sept. 2021), 1–24. https://doi.org/10.1080/15205436.2021.1970186

[75] Brita Ytre-Arne and Hallvard Moe. 2021. Folk theories of algorithms: Understanding digital irritation. *Media, Culture & Society* 43, 5 (July 2021), 807–824. https://doi.org/10.1177/0163443720972314

[76] Andrew Zolides. 2021. Gender moderation and moderating gender: Sexual content policies in Twitch's community guidelines. *New Media & Society* 23, 10 (Oct. 2021), 2999–3015. https://doi.org/10.1177/1461444820942483