"How Do You Quantify How Racist Something Is?": Color-Blind Moderation in Decentralized Governance

QUNFANG WU, University of North Carolina at Chapel Hill, USA
BRYAN SEMAAN, University of Colorado Boulder, USA

Volunteer moderators serve as gatekeepers for problematic content, such as racism and other forms of hate speech, on digital platforms. Prior studies have reported volunteer moderators' diverse roles in different governance models, highlighting the tensions between moderators and other stakeholders (e.g., administrative teams and users). Building upon prior research, this paper focuses on how volunteer moderators moderate racist content and how a platform's governance influences these practices. To understand how moderators deal with racist content, we conducted in-depth interviews with 13 moderators from city subreddits on Reddit. We found that moderators heavily relied on AutoMod to regulate racist content and racist user accounts. However, content that was crafted through covert racism and "color-blind" racial frames was not addressed well. We attributed these challenges in moderating racist content to (1) moderators' concerns of power corruption, (2) arbitrary moderator team structures, and (3) evolving forms of covert racism. Our results demonstrate that decentralized governance on Reddit could not support local efforts to regulate color-blind racism. Finally, we discuss the conceptual and practical ways to disrupt color-blind moderation.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Social and professional topics** → **Race and ethnicity**.

Additional Key Words and Phrases: Platformed racism; color-blind racism; new racism; moderation; volunteer moderator; governance; Reddit; interview

ACM Reference Format:
Qunfang Wu and Bryan Semaan. 2023. "How Do You Quantify How Racist Something Is?": Color-Blind Moderation in Decentralized Governance. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 239 (October 2023), 27 pages. https://doi.org/10.1145/3610030

1 INTRODUCTION

To address harmful behaviors (e.g., hate speech and harassment) in online community spaces, sociotechnical systems of governance are often deployed to structure and organize the myriad user contributions being made [33]. In this paper, governance refers to the evolving structures that mediate and shape the everyday interactions and experiences of those who participate in online spaces [92]. These structures can include everything from policies established by the founders of a given online community to the diverse models through which content is moderated on platforms. For example, organizations like Facebook and Twitter often deploy commercial content moderation whereby paid content moderators engage in hierarchical moderation practices [24, 78]. Other spaces, such as Reddit, rely on a combination of systems, from platform policies to the work of bottom-up volunteer moderators, to govern content and thus the experiences of those who use the platform [19].

Authors' addresses: Qunfang Wu, qunfang.wu@unc.edu, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 27599; Bryan Semaan, bryan.semaan@colorado.edu, University of Colorado Boulder, Boulder, CO, USA, 80309.

2573-0142/2023/10-ART239
https://doi.org/10.1145/3610030

In this work, we focus explicitly on the primary form of governance often instituted in online community spaces and virtual worlds—content moderation. Content moderation is a mechanism of governance where user behavior (e.g., user-generated content) is established as acceptable or not [33]. What is acceptable or not is highly contextual, meaning that determination depends on a range of factors, such as the features of any given platform, how rules are developed and deployed, and the people who are actually doing the work of moderating content or developing rules and norms. How and what content is moderated can shape the experiences of the diverse range of people who participate in online communities. In fact, scholars have characterized how online communities exhibit the characteristics of platformed racism—they enable the construction and propagation of racist ideologies and narratives (e.g., Twitter [17, 70], Facebook [49, 52, 61, 70], Reddit [59, 105], YouTube [67] and others [34, 58]). Online platforms normalize racist speech and grant legitimacy to racist behaviors [52, 61]. Racist narratives thus shape the culture of platforms and marginalize vulnerable groups or push them to the margins of communities [94]. Yet, less attention has been paid to how content moderation might be contributing to this phenomenon.

Given the rampant racism and other problematic content that continues to marginalize people online, there is great interest amongst governments and organizations to develop appropriate frameworks to better regulate behaviors in online platforms [90]. Whereas several exist, content moderation has become the de facto feature mediating people's experiences in online community spaces and other digital platforms [31, 68]. Content moderators have been conceptualized as the "custodians of the Internet"; they serve as a buffer between those who use a platform and the potentially harmful and problematic content being produced by other users that they may encounter, such as hate speech and harassment [31]. Yet, efforts to moderate racist content have been problematic. Prior work has shown that algorithms and content moderation practices are political—they are shaped by societal norms, as well as the individual's or company's politics within which they are designed [31]. In this way, the various technological artifacts, rules, and norms established to govern a platform, and the people involved with making governance decisions, can be biased and reinforce existing social norms that perpetuate racism. On the one hand, various platforms have developed and/or deployed algorithms and tools to identify and moderate racist content. However, research has found that the algorithms and tools used to moderate racist content were in fact most critical of the content being produced by People of Color [79]. On the other hand, on platforms that adopt volunteer human moderation, such as Facebook Groups, Reddit, and NextDoor, moderators have employed a more hands-off approach to moderating racist content, which has, in turn, empowered racist content creators [16, 30, 43, 68]. Taken together, this leads to the question: *how do moderators manage conversations around race and racism in online platforms?*

Our work seeks to answer this question through an explicit focus on how content moderators moderate racist content and how they moderate this content as influenced by the governance—rules and norms—employed by the platforms for which they moderate. To this end, we report on a study investigating volunteer human moderation of racism on online platforms. We build on prior scholarship on governance models in online communities (e.g., [27, 38]) and volunteer moderation (e.g., [64, 87, 104]) in the HCI community. We draw on Bonilla-Silva's conceptual framework of color-blind racism [7] to explore content moderation practices. This study explores this phenomenon through an investigation of content moderation on Reddit–our research site. Reddit adopts both an automatical moderation system[1] and volunteer human moderators to regulate its subcommunities (i.e., subreddits). We recruited volunteer moderators from city subreddits for a semi-structured interview study. Through in-depth interviews with 13 moderators from 11 city subreddits, we obtained rich descriptions of moderators' experiences and challenges in moderating

[1]The automatic moderation system is known as AutoModerator or AutoMod. https://www.reddit.com/wiki/automoderator

racism. Participants reported various forms of racism cases in their subreddits. The results revealed that moderators heavily relied on AutoMod to regulate racist content and accounts. However, covert racism, mediated by and through color-blind[2] ideologies, was typically not addressed. We argue that both human and machine moderation practices are profoundly shaped by and through color-blind perspectives, as illustrated through the challenges of fear of power corruption, arbitrary moderator team structures, and evolving covert racism. We reflect on the findings with prior moderation literature, propose a moderation justice framework, and discuss design implications to disrupt color-blind moderation.

2 RELATED WORK

To address how moderators manage conversations around race and racism in online platforms, we first review online community governance models and discuss the relationship between governance models and content moderation. Here, we argue that human moderators are an essential element in online community governance. Then, on the micro-level, we review how moderators perceived their roles in community governance, highlighting issues regarding power, authority, and responsibilities shared by moderators and other governance actors (e.g., platform owners and users). In the end, we situate the work of moderators in the context of the regulation of racist content. We draw on Bonilla-Silva's color-blind racism and define "color-blind moderation" as moderation mechanisms and practices that are mediated through color-blind racial ideologies. We also provide a review of studies on color-blind moderation and governance in existing literature.

2.1 Governance Models in Online Communities

Governance, broadly construed, refers to the systems by which organizations are run and the means by which constituents are held accountable and responsible for their everyday actions and interactions within a system [50]. It is often invoked to analyze the structures in place to organize and shape behavior within nations, schools, and businesses [45]. The models of governance can vary greatly, depending on the context. For instance, nations have different models of governance through which their people are governed, while businesses often have different structures to organize their members' activities.

In the context of the Internet, scholars exploring online community spaces have examined them through the lens of governance to understand the evolving structures that mediate and shape the behaviors and experiences of their users [92]. Online communities such as Facebook and Reddit adopt different governance frameworks or models to regulate content and user behaviors. Drawing from political science and communication studies, scholars have explored taxonomies and models of governance in online contexts. There are two primary models of governance, based on the amount of agency allocated to different constituents, such as designers and users, within an online community space: the industrial governance model and the community-reliant governance model [38]. The former is a centralized governance model that relies on global policy regulation through top-down efforts by platform creators, while the latter is a decentralized governance model, allocating more agency and power to local volunteers (e.g., sub-community moderators) in shaping and enforcing platform rules [38]. Jhaver et al. highlighted the benefits of community-reliant governance, as it allows for local users' needs to be better served than with industrial governance. Seering further conceptualized this relationship within the context of content moderation, distinguishing between the "platforms and policies" approach and the "communities" approach [84]. The former focuses on top-down governance, while the latter relies on the involvement of sub-community members in the

[2]We note that the term "color-blind" is ableist. We use it here because it is defined as such in the literature. We use "color-blind" and new racism interchangeably.

moderation process. Gorwa proposed a third governance model, known as "co-governance", which goes beyond the frameworks and models that have become prominent in the field [32]. This model suggests the potential of including external stakeholders in the governance of online platforms, even when they are not actual designers or users of the system [32]. Co-governance is composed of self-governance layers, such as platform companies, and users and external governance layers, such as governments, other political actors, and investigative journalists [32].

With regard to the governance frameworks employed by online communities, online spaces have broadly exhibiting an "implicit feudalism" approach whereby (1) power is inherited from platform owners, (2) power is in the hands of a small group of individuals, and (3) the decision-making process is opaque [82]. This is seen in everything from early networked communities (e.g., BBSes and email lists) and peer production communities (e.g., Wikipedia) to current global online communities (e.g., Facebook and Reddit), where the power and authority mediating the behaviors within online communities are held by a small minority of people [82]. As such, scholars have sought to explore and understand the most democratic and effective ways of distributing power and authority in online communities [27, 28, 38, 87]. This has been examined through the lenses of centralized governance and decentralized governance, investigating how much democratic accountability online communities require of their constituents. As of yet, there is no agreed-upon "best" or "most effective" model for online community governance.

On the one hand, scholars have argued for decentralized governance for online communities [28, 38, 87]. Forte et al. observed how online communities often move toward decentralization when they become larger [27]. Through their exploration of local governance cells, such as on WikiProjects, they found that decentralized governance was an emergent property growing out of the need for consensus driven decision-making by local participants. This work highlights the importance of including and empowering the voices of those participating in localized efforts, as opposed to allowing global voices to mediate governance [27]. To discover more feasible governance models, scholars have started to discuss the innovation of decentralized governance models in online communities [38, 83, 106]. Jhaver et al. proposed "multi-level governance," defined as "decentralization in the form of authoritative decision making dispersed vertically across multiple levels, and horizontally over many local governance units" [38]. They argued that platforms with multiple-level governance are more likely to see to the welfare of all users [38].

On the other hand, some researchers have argued that oligarchy is necessary for the survival of online communities [89, 107]. For example, Shaw et al. found that Wiki's governance became more oligarchic as the scale of peer production increased [89]. Centralized moderation was conducted by qualified moderators and, thus, was more consistent than distributed moderation [41]. However, Chandrasekharan argued that oligarchy or centralized moderation is less responsive to public concerns and inquiries and is flawed when it comes to procedural justice [13].

2.2 The Roles of Moderators in Online Community Governance

Scholars have found that both centralized and decentralized governance can benefit online communities; however, these systems of governance can also create and perpetuate structures that harm the most vulnerable members. Distributed governance can lead to the construction and propagation of dominant normative logics shaped by the majority, which can be detrimental to minoritized members of online community spaces [22, 71]. On the other hand, oligarchic governance can marginalize minoritized members through one-size-fits-all policies that overlook the explicit needs and experiences of some users [15, 36]. Geiger used blockbots as an example of how marginalized groups had conducted bottom-up infrastructure reconfigurations to confront harassment [28].

To address harm, online communities have implemented moderation of community content as a governance mechanism to "facilitate cooperation and prevent abuse" [33]. Platform governance

determines the moderation mechanisms [41]. According to Jhaver et al., middle-level governance agents such as moderators played a vital role in coordinating between top-level administrators and bottom-level users in a decentralized governance model [38]. Examples of middle-level governors included moderators on Reddit, streamers on Twitch, and moderators in Facebook Groups, who both enforced platform policies and established community-based norms.

Studies show that moderators perceive themselves to play diverse roles when moderating practices and actions under the different infrastructure and regulatory frameworks of platforms [85, 87, 104]. Moderators in the distributed moderation model exercise more power than those in the centralized moderation model. For instance, on Reddit, volunteer moderators regulated individual subreddits [41]. Since moderators could promote posts, remove content, and ban users, they possessed the ability to profoundly influence and shape the values and norms of their communities [33]. Moderators are seen as civic leaders who regulated and governed their online communities [44]. For example, moderators in the Asian American and Pacific Islander communities developed collective and cultural narratives to confront identity erasure and colonial oppression [21], while other moderators combated a surge of newcomers through active team coordination and a shared sense of community [46]. With more authority than other users on Twitch, moderators were more likely to be imitated, which helped them curate the culture of their channels [86].

In some cases, moderators are given limited administrative power for moderation [64, 65]. Matias described volunteer moderators as "civic labor" who negotiated their authority and responsibility with platform operators, peer moderators, and community users in everyday practices [64]. Moderators on Reddit participated in a blackout to demand platform operators' support and more moderation tools [63]. However, their power did not extend beyond the platform [41]. Sometimes, moderators' authority or power is challenged by community members. Jhaver et al. pointed out conflicts between moderators and users in multi-level governance communities [38]. On Reddit, users were allowed to vote for content, which gave them more power to regulate their communities and indicated that moderators had limited power [93].

Moreover, human moderation work has been viewed as digital labor [66, 98] or emotional labor [19, 75, 104]. Schneider pointed out that the governance ideology of implicit feudalism granted authority to moderators by treating them as a form of uncompensated outsourcing [82]. Moderators were responsible for guarding against harmful, user-generated information such as racist, homophobic, and misogynist imagery, language, or violence on social media platforms [76]. During the moderation process, they needed to acquire background information, fully understood the meaning of an incident, and made a fair judgment, which could lead to mental health issues [75]. Additionally, moderators must deal with complex interpersonal relationships with peer moderators and community members [104].

Moderators' roles are also impacted by how they collaborate with automated content moderation, which employs automated detection and filter tools to identify, prioritize, remove, and ban certain content and/or users. These automated tools provide support to human moderators and users [77], such as automatically removing a significant amount of inappropriate content and providing explanations to users whose content is being removed [36]. The work would be very time-consuming and require emotional labor if conducted by human moderators. However, due to their poor reliability and accuracy in detecting certain types of language, automated content moderation systems are not able to fully replace human moderators [3]. Hence, platforms have commonly adopted hybrid content moderation, which combines both manual and automated approaches [33]. In the hybrid mode, automated tools flag problematic content, and human moderators take a second look [74]. This approach can balance accuracy and productivity by applying both approaches to decision-making [74].

Although they play a critical role, the practices of volunteer moderators in moderating problematic content are still understudied by the HCI community [47]. Scholars have studied the moderation of problematic content in online communities, focusing on harassment [4, 5], hate speech [12], and commercial content [76], and have attempted to develop better tools to fix myriad issues [9, 14, 28, 36, 39, 92]. Our work builds on existing scholarship through an explicit focus on the moderation of racist content and behavior in online communities. Studies have shown that online communities are permeated by racist ideologies [29, 99], yet we do not understand very well how content moderators are working to address (or are perhaps contributing to) the racist logics that have come to dominate online spaces, and thus shape the experiences of minoritized users of online platforms.

2.3 Color-Blind Moderation in Online Communities

More broadly, we know that technologies such as facial recognition [69], judicial [97], and hiring decisions [2] are deeply problematic and shaped by color-blindness. This is best elaborated by Ruha Benjamin, who explains how "the danger of New Jim Code impartiality is the neglect of ongoing inequity perpetuated by color-blind designs. In this context, algorithms may not be just a veneer that covers historical fault lines. They also seem to be streamlining discrimination—making it easier to sift, sort, and justify why tomorrow's workforce continues to be racially stratified. Algorithmic neutrality reproduces algorithmically sustained discrimination" [2]. This shows how technology can be conceived as "race-blind" in how it re-enforces the dominant structure of power and the historical oppression of People of Color on online platforms.

Content moderators, moderation algorithms, and moderation tools have been utilized to detect and combat racist content on online communities and social media platforms. In our work, we focus on the technology of moderation, comprising the people and tools who govern the content and, thus, the experiences of the constituents of online communities. We are interested in how moderation systems and practices are created through or mediated by color-blind racial ideologies—what we dub *color-blind moderation*. Bonilla-Silva found that color-blindness is a new type of racism that has developed and evolved since the Civil Rights era [6]. Color-blindness is the dominant racial ideology in the contemporary US, whereby people use color-blind racial frames, styles, and stories to justify or deny racial inequality at all levels [6, 7]. This ideology borrows from the notion of liberalism and justifies racial matters with non-racial explanations [7]. Thus, color-blind moderation is defined as moderation that is shaped by and through color-blind racial frames. For example, when discussing residential segregation, users and moderators may regard discussions of racism as race-baiting behavior, arguing that segregation is not a matter of racism anymore and that everyone has the right and ability to live where they like.

Recent studies have revealed that moderation on social media platforms does not work as intended and can lead to racially biased moderation. Facebook claimed that its automated moderation system was "race-blind," meaning it treated all races equally [23]. However, Sap et al. found that hate speech detection tools were more likely to flag posts created by African American users or written in African American English as "offensive" content [79]. To address this issue, Facebook shifted to self-regulation, placing the burden of moderation on users [91]. Kelly's study revealed that Black users on Nextdoor were unfairly moderated compared to their White neighbors; for example, Black users were muted by moderators after participating in discussions about race, while inflammatory racist content continued to appear [43]. Reddit applied "quarantine" to isolate certain toxic subreddits, but Chandrasekharan found that this community-wide moderation did not significantly reduce racist language [11]. Matamoros-Fernández noted that platform governance normalized racist humor and abuse on Twitter, Facebook, and YouTube during a racist incident [60].

Previous scholarship has demonstrated that moderation work is steeped in race-blindness and fails to address or reduce racism in online communities. This signals a need to further understand the factors that lead to color-blind moderation. To fill this gap, this study aims to unpack how platform governance and moderation create racial discrimination through color-blind ideologies. To do so, we draw on Reddit as the research site since Reddit applies a hybrid moderation framework of automated and human moderation. In the following section, we will briefly introduce governance and moderation on Reddit.

3 RESEARCH SITE: MODERATION ON REDDIT

Reddit is a social sharing and discussion website that contains many sub-communities called "subreddits." Community members can post content such as text, links, pictures, and videos, which are then voted on by other community members. Reddit enforces three levels of rules to regulate user behaviors: Reddit policies, Reddiquette[3], and subreddit rules [26]. Reddit adopts multiple content moderation mechanisms, and its content policy states that "the culture of each community is shaped explicitly, by the community rules enforced by moderators, and implicitly, by the upvotes, downvotes, and discussions of its community members" [72]. In other words, the culture of each community is co-constructed by both moderators and users.

Reddit has volunteer moderators to regulate its subreddits [64]. Studies have found that these volunteer moderators have limited administrative power for removing content and banning users [64, 65]. Prior studies have highlighted the challenges that volunteer moderators on Reddit face, such as responding to users' complaints of censorship [64], welcoming and educating large influxes of newcomers [46], cleaning commercial content [76], and enacting a jurisprudence record and standardizing community norms for community maintenance [65]. Jhaver et al. investigated Reddit users' attitudes toward content removal and found that users requested more transparency after content removal [35, 37].

Moderators of Reddit can adopt multiple automated tools [10]. The most popular tool is AutoMod, which requires human moderators to set up pre-chosen phrases. AutoMod will scan a post for the presence of these phrases and filter it out for human moderators to review or directly remove it [36]. However, AutoMod cannot fully understand human language, and the pre-chosen phrases need to be continually updated by volunteer moderators [10].

Users can moderate content on Reddit through voting posts and comments [53]. This voting mechanism gives users the power to regulate content [48]. Each post and comment will receive a karma score, which is the difference between the number of upvotes and downvotes. A higher karma score will increase the visibility of the post or comment, as Reddit's post-sorting algorithm prioritizes content based on its karma score. Therefore, the higher the karma score, the higher the visibility. The karma system is a means of content moderation on Reddit [55, 56].

4 METHOD

To understand moderators' views and experiences moderating racist behaviors on Reddit, we conducted a study drawing on semi-structured interviews. The Institutional Review Board office approved the study. We recruited participants from local/city subreddits.

Below, we describe the participant recruitment process, the interview protocol, and the procedure for the data analysis.

[3]Reddiquette is "an informal expression of the values of many redditors" and written by users rather than site operators. https://www.reddithelp.com/hc/en-us/articles/205926439

Table 1. Participants' demographic information

No.	Gender	Age	Race/Ethnicity	Education	Region of Subreddit
P01	Male	40	White	Graduate	West
P02	Female	37	White	Bachelor's degree	West
P03	Male	31	Hispanic	College	Southeast
P04	Female	31	Caucasian	High school	Southeast
P05	Male	51	Caucasian	College	Northeast
P06	Male	54	White/Palawan	College	West
P07	Male	45	White	Bachelor's degree	Northeast
P08	Male	42	White	Graduate	Southeast
P09	Male	51	Caucasian/White	Bachelor's degree	Southeast
P10	Male	25	White	Master's degree	Southeast
P11	Male	33	White	Bachelor's degree	Southeast
P12	Male	44	White	College	Southwest
P13	Female	32	Latinx/Culturally American	Master's degree	Southwest

4.1 Participant Recruitment

We targeted city/local subreddits on Reddit as a research site, as these provide a public sphere for users to discuss issues relevant to their everyday life. As people engage in these conversations, any racism that emerges does so naturally. This strategy of selecting research sites has been adopted in other CHI and CSCW studies of political discourse [51, 88].

The eligibility criterion for recruitment was serving as a moderator or having served as a moderator on Reddit of a subreddit devoted to a city/local area within the US. To identify desirable subreddits, we consulted lists of the most popular city/local subreddits [101, 102]. Popularity was a critical factor, as subreddits with more subscribers were more actively engaged in discussing local affairs. We manually checked the number of subscribers for each city/local subreddit listed in [101, 102] and generated a list of 58 city/local subreddits. Notably, we excluded city subreddits from other countries, since differing histories among countries might lead to incommensurable conversations about race and racism. Thus, our study focused on city/local subreddits based in the US.

We recruited 13 moderators in total, from 11 city/local subreddits, using the Reddit messaging system. This enabled us to quickly and easily contact moderators and send them recruitment ads. Table 1 presents the participants' demographic information; only the subreddit region information is reported to protect their privacy.

4.2 Interview Procedure

Interviews were conducted from July to September 2021 using Zoom and the mobile phone, based on the preferences of any given participant.

We started the interviews by asking some icebreaker questions. We asked participants about their general moderating experiences on Reddit, such as "Can you please tell me the story of why you chose to become a moderator?" and "Can you briefly describe your routine for moderator work in the subreddit(s)?"

Then, we delved into participants' experiences with moderating problematic content. We asked questions such as, "When was the last time you moderated something you considered problematic?" and "Have you had experiences moderating problematic content that was geared toward people's race/ethnicity, gender, sexual orientation, or class?"

If the participants did not mention race-related problematic content, we asked the question, "Have you seen any content that you would characterize as racist?" We also probed further with,

"Can you please give me an example?" "Why do you think this kind of content is racist?" and "Is it hard to identify this kind of content? What makes it hard to identify?" We then asked, "How do you moderate the content containing racism?" and probed further with questions such as, "What are the rules and norms reflecting this kind of racism moderation?" and "Is it difficult to moderate this kind of content and what makes it difficult?"

To collect data on the participants' online experiences with racism and stereotypes in other online spaces, we asked, "Have you ever seen or interacted with racist content in other online spaces (e.g., Facebook, Twitter, Instagram)?" "As regards the racist content on Reddit and other online space(s) you mentioned, are they different?" "What makes them different?"

Lastly, we asked the participants for suggestions for managing racist behaviors and stereotyping online, for example, "Based on your experiences, do you think we should do something more with racist content online?" At the end of the interview, we requested demographic information from the participants (i.e., gender, age, racial/ethnic identity, education, and residence).

After exploring the initial interview data, we revised the interview questions. For example, we added questions regarding the participants' handling of teamwork, since the first few participants significantly mentioned issues with their moderator teams. We asked, "How do you work with your moderator team?" "Do you have any disagreements or conflicts with each other?" and "Are there any challenges with the current moderator team?" We also added a question: "Is there any case of racist content that you did not moderate?" to get participants to reflect on this sort of situation.

During the interviews, we encouraged the participants to share with us some examples of their experiences and describe moderation tools/interfaces, in hopes that this would promote participants' recall and lead to detailed answers.

Each interview lasted approximately one hour. Participation was voluntary and not compensated. All the interviews were audio-recorded with the participants' oral approval.

4.3 Data Analysis

All the interviews were transcribed and combined with our interview notes so that we could glean insights and analyze patterns across cases. The first author conducted an inductive coding process derived from the grounded theory method [95]. This involved coding the first three interview transcripts at the sentence level and generating a codebook. The two authors then met and revised the coding results. Subsequently, the first author coded another ten transcripts independently, adding new codes to the codebook. The codebook reached saturation during this process.

The open coding generated 583 codes in total. The authors discussed the coding results and grouped the codes into 11 categories. The names of the categories, counts of the codes, and examples for each category were as follows:

- Ways/standards used to moderate problematic content on Reddit (236): zero tolerance for racism, dealing with face accounts, different moderation tactics, global rules vs. local rules
- Interactions among different roles (85): moderating in turn, limited feedback from admins, when to provide explanations
- Authority of moderators (28): concerned about power, voting for moderation, example(s) of neutral moderation
- Problematic content/behaviors on Reddit (61): brigading, nomadic racists, cross-posting
- Attitudes toward and suggestions for Reddit design/tools (10): complaints about Reddit tools, disliking new Reddit, the reasons for liking new tools
- General moderation experiences on Reddit (17): subreddit moderating, reasons for being a moderator, time spent on moderation

Table 2. The axial coding results: Themes and their code categories

The authority of moderators
Authority of moderators—Light moderation vs. heavy-handed moderation
Authority of moderators—Being neutral as a moderator
Interactions among different roles—Interactions between Reddit users and moderators
Other platforms—Moderation on other platforms
The governance structures of moderator teams
Interactions among different roles—Relationships and interactions among moderators
Interactions among different roles—Conflicts/disagreements among moderators
The interaction among different roles—Moderator team governance structure
Interactions among different roles—Interactions between admins and moderators
Evolving racist content and behaviors
Problematic content/behaviors on Reddit—Racist content/behaviors on Reddit
Problematic content/behaviors on Reddit—Nomadic racist accounts/subreddits on Reddit
Problematic content/behaviors on Reddit—Covert racist content/behaviors on Reddit
Ways/standards used to moderate problematic content—Challenges moderating racist content
Other platforms—Racist content/behaviors on other platforms
Other platforms—Moderation on other platforms

*: "Light moderation vs. heavy-handed moderation" is a sub-category of "Authority of moderators" (the same below).

- Other issues/facts on Reddit (44): the homogeneous ideology of Reddit, political disagreement, harm of crime-related posts
- Other platforms (59): comparisons with different platforms, Facebook with poor moderation, rampant racism across different sites
- General personal experiences and beliefs (21): the harms of racism to the country, understandings about racism, personal life experiences
- Demographic information (6): age, gender, ethnic identity
- Other (16).

Then we performed axial coding [18] and generated a final coding schema featuring three primary themes: "the authority of moderators," "the governance structure of the moderator team," and "evolving racist content and behaviors." Table 2 shows the codes that were re-grouped under each theme.

Taken together, our analysis found that the perceptions of moderators and their practices were tightly integrated, meaning they mutually informed one another. Specifically, our findings highlighted how color-blindness shaped and mediated moderation systems, from how teams organized their work to how they drew on and used tools like AutoMod. We present are three themes through this lens.

4.4 Researcher Positionality Statement

Researchers' beliefs and values influence the ways they interpret data [96]. We reflect our position and privilege as scholars. The first researcher is an Asian woman who is not an American citizen, and the second researcher is an Iraqi-American, cisgender, heterosexual man who is a member of a

minority, indigenous population in Iraq. To avoid being author colonizers—"much of the work of the academy is to reproduce stories of oppression in its own voice" [100]—the research team conducted open coding of the interview data and carefully developed claims based on the participants' stories. Nevertheless, we need to acknowledge the many intersections of people's race, gender, sexual orientation, socioeconomic status, and other characteristics. Our analysis cannot be representative of everyone's.

Second, the interview interpretation is not related to participants' character or morality. White people may frame color-blind narratives unconsciously [7]. In our work, we identified racism as a problem of power rather than of individual intentions or actions. Our analysis did not aim to identify "good" or "bad" moderators or Reddit users; instead, we focused on understanding the power dynamics of moderation.

4.5 Limitations

For this study, we interviewed the moderators of 11 city/local subreddits on Reddit. The results may not be generalizable to other subreddits or other online spaces.

Moreover, despite our best efforts, most of the participants in our study were White or Caucasian. This biased data sampling reflects the racial and ethnic distribution of moderators on Reddit, which could potentially impact the moderation practices and viewpoints of participants reported in the study. Moderators of Color might adopt different approaches to manage racist content, and this limitation and its impact on content moderation in online spaces are further discussed in the discussion.

The two researchers had no moderation experience on Reddit; therefore, their biases toward moderation work could not be avoided. Virtual ethnography could be conducted in the future, where researchers participate on Reddit as moderators to better understand the contexts and challenges involved.

The interview study did not explicitly ask participants to define what they considered to be racist content in their subreddits. Participants may possess different mental models for this definition, which could be further explored in future work.

5 COLOR-BLIND VOLUNTEER MODERATION IN DECENTRALIZED GOVERNANCE

This study revealed that color-blind racial perspectives mediate human and machine moderation practices on Reddit. Participants provided examples of racist content and behaviors on the subreddits and reported how they applied a range of moderation mechanisms and tools, such as reporting, banning, removal, and the use of AutoMod, to identify and regulate racist interactions. Moderators heavily relied on AutoMod to detect and moderate racist content and accounts, yet covert racism usually remained unresolved. Participants reported challenges of the fear of power corruption, arbitrary moderator team structures, and evolving covert racism when they dealt with racist content. The results are organized around higher-level themes related to how color-blindness mediates moderation. We highlight the complexity of moderation practices, featuring how people's perceptions mediated and shaped their practices and integration of tools into the broader work of moderation.

All participants had experience of moderating racist content. Ten participants shared their experiences of dealing with covert racism. Among them, five participants (P01, P05, P9, P10, P12, and P13) perceived covert racism as coded words or dog whistles. For example, P11 said that racism is rooted in a community's beliefs and attitudes toward Black people, which is not necessarily expressed explicitly but instead manifests itself in the ways people interact with each other and the stances they take on certain topics.

> Yeah. But it's tricky because it's hard to say. It's not like. I've been to 4chan and stuff, like, I know, the kinds of things which get posted, and no one like sharing coded white supremacist messages or dropping in bombs. It's not like that. It kind of stems from these topics that just from growing up in the community, understanding what's at the core of them, and then people fall on either side of it in very predictable ways, which I would guess that if you boil down the fire behind their argument, it comes down to do they like Black people or not? (P11)

To moderate racist content or regulate racist accounts, participants (P01, P02, P03, P04, P05, and P10) reported that AutoMod could effectively detect and remove a wide range of racial slurs and their variants. P10 explained how moderators could create a list of racial slurs to train their AutoMod tools, which would automatically detect and remove them:

> Sure. Um, let's, so you can create, with maybe five minutes of work, a single rule to just remove racial slurs. Like any slur you can think of, you can use regular expressions to make it a little bit more sophisticated, you know, match variations on slurs, you know, replace, you know, the letter L with a one, you know, just to do all of those texts and stuff like that. But you created this very sophisticated matching scheme that can find all of these racial slurs and just remove them as soon as they come [inaudible]. ... They're incredibly useful, incredibly, incredibly useful. Without automated tools, they wouldn't be able to do it. (P10)

However, participants reported several cases of racism that were difficult to moderate using AutoMod or moderators. One such case was when a controversial incident occurred and was posted in their respective subreddits. P06 described managing the influx of posts and comments that followed a mass shooting of African Americans by a Caucasian man in a wealthy area. The shooter, a middle-aged Caucasian man, had been invited to join a group of dark-skinned African Americans at a pool party in La Jolla, but tragically pulled out a gun and started shooting, killing people. In response, people from the condo complex posted updates and videos on their balconies, giving more up-to-date information than the regular news media. This resulted in an influx of news coverage. P06 spent a lot of time managing the influx in the subreddit, manually checking reports, deleting offensive posts, checking users' profiles, and banning racist users.

Besides controversial incidents, participants also reported more complicated racist cases. P03 mentioned a controversial slur that could be seen as disregarding the culture of Hispanic communities. The International Association Football Federation (FIFA) issued a ban on a Mexican chant that had been used for many years. The chant was deemed a gay slur by some communities, but the Hispanic communities did not view it as such. This point of view was echoed by members of the subreddit r/MLS[4], who were against the Hispanic crowd using the chant. However, in other Hispanic communities, it was seen as having a completely different meaning:

> So lately, FIFA has banned Mexican fans from saying a certain chant that they've been saying for many, many years. Because it was a gay slur. But in the Mexican community, there's not a gay slur in my community as well. It's not a gay slur. But it's funny, saying because, in MLS and r/MLS, which is the American subreddit, they're, like, very anti the Hispanic crowd saying this chant. And, but it's funny because you go to, like, other Hispanic communities there just like. No, it just means, you know, means something completely different. (P03)

[4]r/MLS: The central hub for all levels of soccer in the US & Canada, especially Major League Soccer. https://www.reddit.com/r/MLS/

With the reported various understandings of racism and moderation practices, this study took a further step to investigate the challenges and issues in dealing with racism on Reddit. Taken together, we argue that moderation on Reddit is shaped by and through color-blindness. The decentralization of governance on Reddit led to a passive and inconsistent moderation of racism, hindering any local efforts to revise color-blind racism. Three primary challenge themes emerged: 1) the responsibility of moderators, 2) the team structure of moderators, and 3) evolving covert racism.

5.1 The Pitfall of Power Corruption: Limiting Moderators' Responsibility

The interviews revealed that color-blind moderation was partly due to participants' fear of power corruption. The fear limited moderators' responsibility, including their regulation of racist content. A major issue is that White members of the population either remain ignorant of racism or take no action against it. As a result, moderators perpetuated racism through their unwillingness to confront color-blind racism.

Moderators on Reddit wield a great deal of power. They can create their own rules for their subreddits and take whatever moderation actions they deem necessary. Some participants have expressed concern that this power may be abused. P12 used the phrase "power corrupts" to describe moderators who had overstepped their bounds in subreddits. Other participants raised questions about the legitimacy of moderators' power and the complaints from Reddit users. In response to these concerns, some participants reported that they had adopted a light moderation approach. This led to a form of color-blind moderation, where moderators were not proactively monitoring the racist content of the subreddit.

First, one of the concerns about power corruption was that personal beliefs and judgment might intervene in moderation standards, leading to dictatorial moderation. Thus, some participants avoided moderating based on personal beliefs. P04, P05, and P08 expressed that they held different opinions toward problematic behaviors when they were in the roles of users and moderators, with an implicit meaning that some moderation practices are inconsistent with their own beliefs. For example, P04 explained that her moderator team was careful about making moderation decisions regarding problematic content. She tried not to use personal beliefs to "cloud" her judgment and instead wanted to be "objective" and "unbiased." It also became clear that P04 and her team hesitated to deal with content that went against their own beliefs. They were very cautious about being dictators. She used the words *accountable* and *impartial* to describe the importance of being "neutral" in moderation practices:

> The biggest thing that we have is trying to keep our personal beliefs and politics and everything from clouding our judgment. Like, especially with [subreddit name], because our moderation team almost universally leans pretty far to the left by American standards. So we kind of have to take a step back and try to be as objective as we can and not let our own personal bias lead to it. So we have a pretty, pretty clear cut, I guess, kind of our own internal judgment as to whether or not was, like, being a little too biased.
>
> And, like, we always, like, if we're not sure, we always ask each other, like, like, hey, am I right here? Or, like am, am I letting my own beliefs, like, cloud judgment here? I like things like that, and we try to hold ourselves accountable and be as impartial as possible. (P04)

Second, the concerns of power corruption arose because the position of moderator did not grant any authority. Moderators felt that their power was not legitimate to use. For example, P05 explained that moderators were in a position of authority. Nevertheless, he quickly withdrew the

term *authority* and instead used "the police for the subreddit" to describe the role of moderators. This quote suggested that even though P05 did not agree with the idea of policing, someone needed to be empowered to stop troublemakers in order to maintain order and prevent censorship:

> And sometimes, I want to call someone out. And I feel like maybe it's not appropriate. Because, I mean, I guess a position of authority, it's not really much authority, but you know, it's like the police for the subreddit. ... [W]e get other people saying, oh, why was this post deleted? You know. And they complain about that, you know, and they tried to say we're censoring them as if we're the only source of media. ... [I]f someone starts acting anti-socially, it's not really for the general public to act, right. You need someone with authority, or as I'm saying this, I'm surprised because I'm not pro-police at all, you know, but someone needs to be empowered to stop troublemakers. (P05)

P05 and P12 reported complaints from users claiming they felt they were being overly moderated. P12 gave an example of the dilemma they faced when moderating subtle racism, which likely received support from users through upvotes. If the moderators removed the subtle racist content, they would be accused of biased moderation by those who supported it. Conversely, if they left it, they would be accused of the same:

> That subtle kind of racism, people will upvote that. And sometimes it's like, if I remove it, that's really, that's kind of a tough call. Because if I remove it, then I get accused of being biased. And if I don't remove it, then someone's gonna say, why don't you remove this racist crap? (P12)

Given their concerns about power corruption, the participants reported that they would apply light or mild moderation rather than heavy-handed moderation—as P08 described it, "laissez-faire, open and passive in moderation." This open and passive approach, whereby moderators are able to maintain a neutral standpoint and let the community (users) decide what is good and bad by voting, allows for the majority voices, often of White supremacists, to dominate. P13 said her moderation followed the reactions of users—if users reacted lightly to the content, she would do nothing to it; if many users were bothered by the content, she would remove it. P11 used the killing of Breonna Taylor[5] as an example. Breonna Taylor was a 26-year-old African American woman who was shot to death in her apartment when police officers used forced entry for a drug-dealing investigation. Taylor's boyfriend, Kenneth Walker, was with her in the apartment. P11 did not give specific context about how users talked about this incident, but he chose not to ban users from discussing this kind of situation. P11 believed that given the demographics of this subreddit, similar to those of Reddit, users would downvote hate speech:

> But clearly, the implications of if you believe that Breonna Taylor deserved to get shot because you believe her boyfriend was a drug dealer, and it was an honest mistake by the police to shoot her but serve you right for being around a drug dealer? Is that racist? I don't know. I mean, I would presume that's coming from a racist place. And I, you know, will downvote that comment. But is it overtly racist? I think there's room for interpretation. And again, I think that goes into the situation of when to ban and when to downvote. So, in a lot of these situations, I'll just not ban [them], but I'll downvote them, and then everyone else would have, too. And so I know, I know, that community is very, because of the demographics of the community and Reddit in general, are very active in, you know, wanting to force out hate speech online. They do a great job of flagging crazy stuff, you know, so I trust the community to take care of that content. (P11)

[5]Killing of Breonna Taylor. https://en.wikipedia.org/wiki/Killing_of_Breonna_Taylor

Overall, the moderators' concerns about power corruption, to varying degrees, undermined their authority as the guardians of Reddit communities and resulted in hesitant moderation of problematic content, particularly content that included subtle and covert racism. The light moderation led to color-blindness, allowing White supremacists to dominate.

5.2 The Team Structure of Moderators: A Roll of the Dice

Participants discussed how the structure of moderator teams could affect the moderation of racist content. It was noted that the structure of a moderator team is often determined by the supreme moderator's personality or ideology and can lead to power shifts. In a hierarchical structure, if senior moderators are "blind" to race, junior moderators may struggle to challenge the regulations. In a flat structure, moderators may opt not to address subtle and covert racism in order to avoid conflict. Both structures lack arbitration mechanisms, resulting in an unstable performance in moderation and an inability to deal with racist behaviors.

In relation to this, we provide background information on the moderator team on Reddit and report participants' experiences. If a user creates a new subreddit on Reddit, they automatically become the subreddit's moderator. They can then invite other users to join them in governing the community. A list of moderators in chronological order is available in each subreddit. The moderators listed at the top are the senior ones, while those at the bottom are the junior ones. The user who created the subreddit is listed first and is considered the "supreme" moderator. This moderator can assign another moderator to take on this role.

The participants reflected on the moderator teams of their subreddits, which revealed a great variety of team structures. These structures can be broadly categorized as hierarchical and flat. In a hierarchical team, the senior moderator(s) would take charge of other moderators' assignments and make primary decisions (e.g., rule revision); in a flat team, moderators would usually apply the majority vote principle to make decisions.

According to the participants' reflections, the structure of a moderator team is based on an arbitrary decision-making process, determined mainly by the supreme moderator, or by the supreme moderator's personality or ideology. Reddit yields control to moderators and cannot intervene in such situations. This can lead to an oligarchy or unstable performance in moderation, including failure to deal with racist behaviors. Furthermore, it is easy for arbitrary decision-making to result in color-blind moderation, no matter whether the team structure is hierarchical or flat. In a hierarchical moderator team, if the senior moderators are "blind" to race, it becomes difficult for junior moderators to challenge the regulation of rules and norms framed around color-blindness. In a flat moderator team, moderation can be color-blind, too. Moderators may hold nuanced attitudes toward racism, even if they claim to be anti-racist, and to avoid personal differences, they are more likely to compromise over inaction toward subtle and covert racism. Therefore, the structure of a moderator team can influence the moderation of racist content or even re-enforce racist behaviors.

First, in a team with a hierarchical structure, the moderator would not resolve disagreements in a democratic way, which can lead to intense conflicts among moderators. In extreme circumstances, some moderators were removed from the team by the senior moderator(s), or they left the subreddits by themselves. In one case, P05 recalled that in a subreddit he moderated, the first moderator was called the "boss," meaning the team structure was hierarchical. The "boss" performed very hands-off moderation. He had different opinions about dealing with racism and other problematic content. And one day, P05 found that he and several other moderators had been removed by the "boss" from the moderator list. After that, P05 realized that the "boss" was running a moderation experiment to see what would happen in the subreddit with very light moderation. This resulted in blatant racist narratives being widely disseminated on the subreddit. P05 and other moderators reported this to the Reddit administration and were later added back to the list. This experience highlighted how

higher ranking moderators might not take into account racial issues, which can make it difficult for lower ranking moderators to question the rules and standards that ignore race:

> The thing, it wasn't even a disagreement. It was—so the guy, he had views on censorship where he basically felt like it should be completely hands-off. And I didn't agree with that because I didn't want to see this racism and stuff, and neither do most people. ... He was very hands-off, like he would disappear for weeks sometimes. And then he would be popping up with strange rules. ... So I went, I looked at the page, and I saw I had been removed as well. And then I looked at the list, and I saw we were all removed. ... I think he kind of looked at it as an experiment and just wanted to see what would happen. (P05)

P01 and P06 reported similar cases, where moderators in one subreddit would not address disagreements, leading to some moderators leaving and creating a new, equivalent local subreddit. P06 said the new subreddit competed for participants and played the devil's advocate against the first one. P06 was harassed with toxic language in the new subreddit. P01 reported that in the city subreddit he was moderating, the users were divided into two groups—the moderates and the progressives. The progressives are democratic socialists, and the moderates are more traditional Democratic Party candidates. Issues in the city subreddit were often debated around progressive policies versus moderate policies, for example, whether or not to build taller buildings or build more buildings:

> So now we've subdivided sort of like the Democratic Party into the two groups called the moderates and the progressives, and they generally the progressives are democratic socialists, and the moderates are more traditional Democratic Party candidates like very far left of center, but like not actively pushing for like ending the concept of private property and stuff like that. So this, I mean, the progressive versus moderate divide is it has to be significant in the city. Right, because it's first past the post. So all the political candidates are trying to, there will always be a divide somewhere. And so the issues that people fight about in [the city name] are generally related to progressive policies versus moderate policies. (P01)

Moderator teams with a flat structure are better at dealing with controversial behaviors. P10 explained that in a more democratic team, when moderators encountered racism and were unsure how to deal with it, they could benefit from the team discussion, which helped to bolster their confidence and enabled them to make better decisions:

> Because if you're not sure, you know, you can ask the team, "Hey, I think this breaks our rules against racism. This is what I'm thinking of as the punishment for it. What do you think?" You know, other people will in and will come to some sort of consensus, we'll take action, and then it's a team decision, you know, because we can rely on each other to help make better choices as far as moderation policy goes. (P10)

Most participants preferred a flat team structure; however, this did not guarantee that racism would be addressed effectively. P01 noted that he and another moderator in their subreddit were both progressive and had a good relationship, but they had different levels of tolerance when it came to racist narratives. P13 mentioned that sometimes, she and other moderators could not reach an agreement and desired to be open to divergent opinions on racist narratives. The flat team structure could also lead to an inability to effectively address racist behaviors.

In short, in both hierarchical and flat structures, moderators lacked an arbitration process to address racist behaviors. In the hierarchical structure, senior moderators might be blind to race, making it difficult for junior moderators to challenge the regulation of rules and norms that

were framed around color-blindness. In the flat structure, moderators often compromised over inaction toward subtle and covert racism to avoid personal differences. The lack of arbitration mechanisms, such as voting, resulted in an unstable performance in moderation, leading to the failure to effectively address racism.

5.3 Evolving Covert Racism: Sluggish Top-Down Deployment

Most participants reported that the current moderation tools (e.g., AutoMod) were useful for detecting and removing racist content. However, some still faced challenges in dealing with emerging forms and behaviors of covert racism, such as racial slurs, new memes, racist usernames (screen names), and entire racist subreddits. These new forms had nothing explicitly to do with race, but were coded with racial buzzwords. Moreover, there was no way to report a user because of their racist screen name. The participants noted that the top-down deployment of moderation mechanisms and policies was too slow to react to these new forms of covert racism, which condoned color-blind moderation or disregarded evolving covert racism on Reddit.

Several participants reported that it was challenging to figure out if a dog whistle was racist or not. For example, P13 mentioned that sometimes she was unsure if a meme was a racial dog whistle for hate speech or just a weird meme. When P13 tried to moderate this kind of meme, she received pushback from users, asserting that "this is a good meme" and not a dog whistle. P09 described how, in conversations, users turned a slang expression into a dog whistle. In these examples, the users were confident that they could circumvent punishment because they knew color-blind governance still dominated this site.

P05 provided an example of users' using "inner city" to refer to Black neighborhoods. *Inner city* is a racially coded phrase used in the conservative news. To deal with this coded word, moderators had to be prepared for users' objections that "I didn't say anything wrong."

> But you know, when the news when they say inner city, you know, they mean Black people, right? That's, that's a code word for the conservative news to mean, you know, Black people. (P05)

To cope with these ambiguous dog whistles, P09 developed a sensitivity to whether users were exchanging dog whistles and turning conversations racist. He tried to keep up with new memes that were becoming popular on Reddit:

> But it's, it does allow me to kind of keep up on some things that people use in order to start, starting to turn conversations away from something useful into something racist where you don't know. (P09)

Despite praising the usefulness of AutoMod for moderating racism, the participants also highlighted its shortcomings. P10 noted that AutoMod is not yet sophisticated enough to detect subtle or disguised racism in language, which requires a human eye to accurately identify:

> Always no. They can do the simple stuff. And some of you know the more complex, complicated stuff, but you know, identifying racism gets much, much harder as users get more sophisticated. So, you know, we could have a user post a, you know, a four-paragraph essay, trying to explain why Black people commit all of the crimes in the United States, and they wouldn't just straight up say that. They would say, you know, well, according to these government reports, they might even include links, and they might just include dog whistles for racism that aren't, you know, quite as clear-cut as just like, you know, 13% of the population but 50% of the current, you know, biased statistics from the FBI or something like that. So identifying those instances of racism as much as it takes a human eye, we can't really automate that. At least not yet. So things like that we, we do manually. (P10)

P03, P06, and P10 conducted manual checks if AutoMod could not process covert racist narratives. After a comment was filtered out by AutoMod, they investigated further to determine whether it should be removed for the right reasons. As an example, P03 checked the poster's comment history to see if they had a history of making similar comments. This allowed P03 to make an informed decision about whether to ban the poster or not:

> For example, the N-word, any racial slurs that are offensive, you just added to the filter. Anytime someone makes a comment, the filter automatically removes a comment right away without anybody having to do any action. We get a notification every time that happens. ... Make sure to take a look, and so one of the first things that we would do is make sure that it was removed for the right reason. We do a little bit more investigation. The first thing I do, what I would do is read the poster's comment history to see exactly what they were, just to see if they have a history of doing this anywhere else before we banned them or anything like that. (P03)

Racism can also be embedded in users' screennames and subreddit names, making moderation more difficult. For example, P10 explained that there were a lot of anti-Jewish words embedded in usernames. Reddit has forbidden these racist behaviors in usernames or subreddit names in its community policy[6]. However, there were no moderation mechanisms for dealing with these racist usernames or subreddit names. Moderators could remove racist posts or comments, or they could ban a user from posting or commenting, but they could not report someone for their username. As P10 mentioned, this issue remained unaddressed for a long time, frustrating users.

Several participants reported having encountered racist subreddits on Reddit and expressed concerns about the lack of moderation mechanisms to regulate them. For example, with racist usernames, there was no button for users to report the subreddits; instead, it was up to the Reddit administrative team to manage them. However, before the administration team could shut a subreddit down, it could remain active for a long time, creating and disseminating racist content. Even when a racist subreddit was shut down, the users could move to other subreddits or start new ones. For example, P02 found one possibly racist subreddit called r/Europe, with about three million subscribers. She suspected that the current moderators had taken over the subreddit and turned it into a space for White supremacists. P02 also estimated that there were about 1,000 racist subreddits, but she, as a user, could not find them on her own:

> I mean, some like, you know, one subreddit that I think, I think it's r/Europe is, is just a bunch of racists is what I've been told, like a whole bunch of White supremacists, and I don't know how they managed to take over the whole subreddit r/Europe. ... I mean, there are probably 1,000 little subreddits that are racist, but I can't find them all. (P02)

Given the challenges of emerging forms of racism and limited moderation resources, the participants suggested that Reddit should provide more resources to support moderators, such as policy enforcement, moderator training, and racism detection tools. P10 was a very experienced moderator. He had been a moderator for multiple subreddits, four of which had more than 20 million subscribers, and he used to be the head moderator for two subreddits. When asked what suggestions he would give to assess and manage racist behaviors, he pointed out a fundamental question: "How do you quantify how racist something is?" Even with his rich experience moderating both overt and covert racism, he was still troubled by this question, noting that detecting racism requires one's perceptual and rational judgment, and even training to make a judgment:

[6] "Do not threaten, harass, or bully" in Rule 1—"Behavior can be harassing or abusive regardless of whether it occurs in public content (e.g., a post, comment, username, subreddit name, subreddit styling, sidebar materials, etc.) or private messages/chat" from https://www.reddithelp.com/hc/en-us/articles/360043071072

> How do you quantify how racist something is? Is that even a reasonable question to, is that something reasonable to even attempt, like how would you quantify how racist something is, and my personal answer wouldn't be, "I don't think you can kind of make a qualitative judgment, um, that requires, you know, experience and intuition and training and empathy to actually get right." (P10)

Apparently, P10 did not think Reddit provided sufficient resources to support moderators in judging racist behavior. He called for better top-down policies from the Reddit administration. In fact, Reddit and most subreddits forbid racism and provide examples of racism (e.g., "Post describing a racial minority as sub-human and inferior to the racial majority," "Meme declaring that it is sickening that People of Color have the right to vote.") [73]. Nevertheless, P10 argued that it was not the responsibility of the moderators to create rules against racist or other problematic content, although most subreddits had those sorts of rules. Instead, the platform Reddit needs to make anti-racist policies more explicit and enforce them. To ensure these policies are enforced, Reddit could provide training for moderators on how to manage this kind of content and consider their well-being.

To better address the issues of new forms of racism, participants suggested the implementation of moderation tools that could be applied universally across the platform. For example, in regard to concerns about racist subreddits, P02 suggested that Reddit should provide a site-wide list of slurs that could be used to detect comments containing racist content, rather than having subreddits do it individually. This would not only help the Reddit administration identify potentially toxic racist subreddits, but it would also benefit moderators:

> I mean, at the very least, they should just take that list of slurs and put it in, like a site-wide comment capture, like, there's no reason you shouldn't have to do that on an individual subreddit. (P02)

Other participants shared their experiences of using external tools such as tagging tools for labeling racist users, word-cloud tools for illustrating trends, and detection tools for hiding toxic content or estimating its toxicity. All these efforts or experiences demonstrate Reddit's sluggishness in updating its moderation mechanisms to react to evolving forms of racism.

6 DISCUSSION

Racism and other problematic content online have drawn more attention from governments and organizations to develop appropriate regulatory frameworks for online platforms [90]. Content moderation has become the necessary attribute of platform regulation [31, 68]. Studies have revealed that current moderation systems on platforms have resulted in racially unfair moderation [11, 43, 60, 79]. This study deepens the understanding of the current moderation scholarship by exploring how color-blind racial ideologies mediate moderation mechanisms and practices on Reddit, which we dub *color-blind moderation*. Color-blind moderation is exhibited through the laissez-faire moderation of racism, arbitrary team structures inclining to color-blindness, and sluggish top-down deployment that fails to keep up with evolving covert forms of racism. We reflect on the findings with prior moderation literature, propose a moderation justice framework, and discuss design implications to disrupt color-blind moderation and support moderators' work.

6.1 Reflection on Color-Blind Moderation Practices

The moderation practices on Reddit are profoundly shaped by and through color-blind racial perspectives. Prior studies have demonstrated that color-blindness is embedded in the moderation practices of platforms [11, 43, 91]. This is especially true on Reddit, where the majority of the participants in this study were White or Caucasian, reflecting the overall demographic distribution

of Reddit, which leans toward White. In 2016, about 70% of Reddit users were White non-Hispanic in the US [81]. This dominant White culture on Reddit has been perpetuated in moderation practices, including the moderation work of AutoMod, the burden of race taxation in volunteer moderation, and users' voting and reporting on racism.

6.1.1 Color-Blind Moderation with AutoMod. This study revealed that moderators heavily relied on AutoMod to detect and moderate racist content and accounts. However, covert racism, framed through color-blind ideologies, was typically not addressed by the system. Automated moderation systems have often been criticized for their one-size-fits-all approach to different types of hate speech [90]. Existing linguistic frameworks have not been able to capture the implicit implications where people express social differences and power imbalances through language [80]. Racism has usually been treated as a single kind of hate speech [62], making it difficult for technology to identify implicit inferences associated with covert racism [8]. As a result, automated systems are unable to detect and address this type of racism.

To better address the issue of race-related content, research suggests that there is a discrepancy in how White and non-White raters process tweets with racial topics [54]. Specifically, non-White raters may view certain tweets as negative, while White raters may view them as harmless, positive, or neutral [54]. This discrepancy could explain why many covert acts of racism were not detected or regulated by moderators or automated tools. White moderators may not be as sensitive to (or may be ignorant of) the subtleties of racism as non-White moderators on Reddit. Additionally, automated hate speech detection tools have been found to be more offensive to content created by People of Color than content created by White people, suggesting that automated tools may not be as race-blind as previously thought [23, 79]. This study did not find unequal automated detection; however, future work should focus on the views of non-White moderators or users.

Our results showed that racism is widely considered a norm violation across Reddit and is written in the rules of many subreddits. However, the definition of racism is not explicitly addressed in these rules. Chandrasekharan and colleagues [13] found that hate speech associated with racism and homophobia is widely regarded as a norm violation across Reddit; they define a norm as a "hidden rule" that can guide moderators when they create rules for their subreddits [13]. Fiesler and colleagues [26] further coded the rules of 100,000 subreddits, where hate speech was addressed in the main rules, but racism was not explicitly defined. This highlights the importance of understanding racism and its implications, as well as providing resources to help moderators address it more effectively. To ensure that racism is adequately regulated, it is essential to explicitly define racism in the rules.

6.1.2 "Race Taxation" in Volunteer Moderation Work. Scholars have coined the term "race taxation" or "cultural taxation" to refer to the additional burdens faced by People of Color in their work or other obligations [42]. Similarly, in our research, we found that moderators were subjected to various forms of "race taxation" when dealing with racism in their moderation work. This type of labor has been labeled as digital labor [66, 98] or emotional labor [19, 75, 104], as moderators are expected to take on an emotional burden while working for free, which can be taxing for them.

Our study revealed that moderators needed to address the evolving forms of covert racism. Roberts et al. found that moderators served as shields against harmful user-generated information such as racism, homophobia, and misogyny on social media platforms [76]. During the moderation process, they need to acquire background knowledge, fully understand the meaning of an incident, and make a fair judgment, which can lead to mental health issues [75]. Jiang et al.'s research revealed that moderators faced new challenges, such as detecting disruptive noise while moderating voice-based online communities; they must develop strategies to deal with the ephemeral nature of voice communication [40]. These challenges could result in ineffective moderation or false

accusations [40]. Additionally, as Reddit does not provide a clear definition of racism, moderators must provide one and educate Reddit users about racial slurs, leading to extra work if AutoMod mistakenly "catches" posts with language that is not explicitly racist [87].

In addition, moderators need to manage complex interpersonal relationships with their peers and community members [104]. Weld et al. found a discrepancy in perceived democracy between moderators and users on Reddit [103], and there may be personal conflicts between moderation teams when different opinions are expressed on racism [20]. Moderators must navigate a fine line between their power and responsibility when dealing with racism. If they choose to moderate aggressively, they could be accused of being too harsh or of abusing their power. In our study, several moderators reported being harassed or removed from their teams due to personal conflicts.

Moreover, Reddit heavily relies on volunteer moderators who are not compensated for their work [82]. Schneider argued that this lack of payment and recognition for the responsibility that moderators bear, often for their own subreddits, is due to the platform's governance ideology of implicit feudalism, which grants authority to moderators through uncompensated outsourcing [82]. Furthermore, moderators are unable to negotiate collectively with platform administrators about their role due to not being employees, which adds to their burden.

6.1.3 Users' Voting and Reporting on Racism. In online platforms, users often participate in moderating racist content through voting and reporting mechanisms. Our results showed that when moderators were concerned about power corruption, they preferred community members (users) to report or vote against harmful content or accounts. Community members leveraged voting and reporting mechanisms to achieve their collective goals. For example, Reddit users integrated authentic information and carefully used upvotes to increase the visibility of crucial information during crisis events [55, 56].

User voting and flagging mechanisms have been proposed as a way to moderate hate speech and racism online [28]. Geiger et al.'s research indicated that collective, bottom-up technical mechanisms could be effective in addressing racism in online community platforms. These mechanisms involved the development of blocklists and users choosing to subscribe to them based on their preferences. While this approach may be beneficial for marginalized groups, user voting and flagging mechanisms can be manipulated to amplify certain viewpoints and censor others [105]. This was confirmed by our study participants, who reported that their experiences with user voting and flagging mechanisms in some subreddits had resulted in the reinforcement of racist culture.

6.2 A Race-Conscious Justice Framework for Online Platforms

This work reveals the many challenges of moderating racist content and user behaviors on Reddit. These challenges relate to how moderators use their power, how they structure their teams, and how the platform and moderators react to new covert racism. All of these issues point to the larger issue of how platforms can eliminate color-blindness through governance and what roles moderators can play.

We suggest that platforms develop a "race consciousness" to disrupt color-blind moderation. According to Bell, race-consciousness means "to be awake, aware, mindful, informed and intentional about challenging racism and working toward racial justice" [1]. Such a racial-conscious lens can help to "unmask apparently nonracial phenomena as precisely racial in nature" [57]. Race is still a salient factor in many online interactions, and a race-conscious lens confronts ignorance about issues of race and normalizing hierarchies of power and privilege. Bell [1] suggested that rather than focusing on intent, it is more important to focus on the outcomes of practices and policies, as they can produce racism and harm People of Color, even when not intentional. Thus, moderation should focus on the outcomes of their moderation rather than guessing at the intent of users.

To address the issue of moderating racist content and user behaviors, we propose a race-conscious justice framework, which is tri-leveled comprising the levels of governance, moderators, and users. This justice framework seeks to ensure that platforms are held accountable for governing racism, that moderators have the necessary tools and resources to moderate racist content and user behaviors on Reddit, and that users have the reporting mechanisms to help moderate racism. At the governance level, the framework seeks to ensure that data and policies do not reinforce existing power structures and inequalities and make platforms more transparent about how policies and tools (e.g., AutoMod) are designed. To accomplish this, platforms need to make their policies and tools more race-conscious. Additionally, they should provide users and moderators with training in racial literacy. At the moderator level, the framework seeks to ensure that moderators have the authority and resources to effectively moderate racist content and user behaviors on Reddit. Moderators can seek and implement both internal and external tools to regulate their communities. At the user level, the framework seeks to ensure that users have the effective mechanisms to vote and report racist content and user behaviors.

6.3 Design Implications

In light of the findings and the proposed moderation justice framework, we discuss design implications to dismantle color-blind moderation.

Moderation of platforms needs to refer to the experiences and knowledge of People of Color who have been exploited on platforms [54, 90]. To this end, Larimore et al. proposed that annotations of anti-Black racism should be based on the interpretation of Black people [54]. To better recognize what is implied by covert words, scholars have proposed and developed state-of-the-art practices. For example, Sap et al. introduced the *SOCIAL BIAS FRAMES* to model the pragmatic frames in which social inequalities and biases are projected onto others and created the Social Bias Inference Corpus [80]. Elsherief et al. developed a theoretical taxonomy of implicit hate speech and created an evaluation dataset with precise tags [25]. These methods and datasets have incorporated the perspectives of People of Color in order to mitigate implicit linguistic biases and stereotypes enabled or reinforced by language models. This will help to develop moderation tools that can detect the subtle nuances of covert racism and other biases.

Moderation practices can help address biased reporting or voting systems by creating more diverse and inclusive moderation teams. To ensure racism and other forms of discriminatory content are identified and addressed, moderators should be composed of people from different genders, races, and ethnicities. Furthermore, moderators should be provided with training and education in recognizing and responding to racism and other forms of discrimination. Additionally, Reddit should provide financial and emotional support to moderators, such as compensation and mental health resources. To encourage users to report racism and other forms of discrimination, the platform should also implement clear and effective mechanisms, such as elections, boards, juries, and petitions, to enable more democratic decision-making and ensure greater accountability for moderation teams.

Moderation tools can be created in one community and easily embedded in the governance infrastructure of another community. To facilitate this, platform owners need to build a fundamental governance infrastructure. For example, on Reddit, many moderation-related subreddits (e.g., r/ModSupport[7]), third-party tools, and blocklists have been created to enable moderators to better manage their communities. Channel chat moderation modes are tools available to limit how users post in chatrooms; they were found to reduce spam across all three modes [86]. However, these resources have not yet been adopted across the site nor tailored to target different types

[7]Mod Support. https://www.reddit.com/r/ModSupport/wiki/moderator-tools/

of behaviors. Moderators have to navigate these resources on their own. To speed up innovation deployment, platform owners can integrate resources and distribute them to moderators. The moderator participants reported that they could test AutoMod configurations for racial slurs and run experiments for different moderation ideologies. Moderators supported maintaining a smooth flow of innovation. To promote the innovation of governance and moderation, the platform needs to set up an innovation pipeline that allows moderators to test and report innovative practices and tools [83]. This system enables moderators to experiment with modular innovations and improve their moderation strategies.

7 CONCLUSION

This study reveals color-blind content moderation practices on Reddit. AutoMod was found to be cumbersome to detect covert racism, especially when framed as color-blindness. Additionally, moderators were encumbered by other aspects of the decentralized governance structure, such as power corruption, arbitrary team structures, and evolving covert racism. It is evident that there is a need to rethink the efficacy of Reddit's decentralized governance model in addressing racial oppression across the site. Our research contributes to scholarship in pushing back against racism and racist ideologies in online spaces, and we hope that it will help to create meaningful change.

ACKNOWLEDGMENTS

We are deeply grateful to the National Science Foundation under grant no. 1657429, and the Doctoral Students Awarded Summer Dissertation Fellowships and Dissertation Fellowship at Syracuse University for their support. We also wish to thank the reviewers for their invaluable feedback on our manuscript. Lastly, we extend our sincere appreciation to all the participants who kindly provided their time and effort to this project.

REFERENCES

[1] Lee Anne Bell. 2016. Telling on racism: Developing a race-conscious agenda. (2016).
[2] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social forces* (2019).
[3] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics*. Springer, 405–415.
[4] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When online harassment is perceived as justified. In *Twelfth International AAAI Conference on Web and Social Media*.
[5] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.
[6] Eduardo Bonilla-Silva. 2016. Down the rabbit hole: Color-blind racism in Obamerica. *The myth of racial color blindness: Manifestations, dynamics, and impact* (2016), 25–38.
[7] Eduardo Bonilla-Silva. 2018. *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers.
[8] Mondher Bouazizi and Tomoaki Otsuki Ohtsuki. 2016. A pattern-based approach for sarcasm detection on twitter. *IEEE Access* 4 (2016), 5477–5488.
[9] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. 2017. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3213–3226.
[10] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.
[11] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.
[12] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the*

ACM on Human-Computer Interaction 1, CSCW (2017), 1–22.
[13] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.
[14] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3175–3187.
[15] Angshuman Choudhury. 2020. How Facebook Is Complicit in Myanmar's Attacks on Minorities. https://thediplomat.com/2020/08/how-facebook-is-complicit-in-myanmars-attacks-on-minorities/. *The Diplomat* (2020).
[16] Andrew R. Chow. 2022. Reddit Allows Hate Speech to Flourish in Its Global Forums, Moderators Say. https://time.com/6121915/reddit-international-hate-speech/. Accessed by 04/11/2022.
[17] J David Cisneros and Thomas K Nakayama. 2015. New media, old racisms: Twitter, Miss America, and cultural logics of race. *Journal of International and Intercultural Communication* 8, 2 (2015), 108–127.
[18] John W Creswell and Cheryl N Poth. 2016. *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.
[19] Bryan Dosono and Bryan Semaan. 2018. Identity work as deliberation: AAPI political discourse in the 2016 US Presidential Election. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[20] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[21] Bryan Dosono and Bryan Semaan. 2020. Decolonizing Tactics as Collective Resilience: Identity Work of AAPI Communities on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–20.
[22] Stefanie Duguay, Jean Burgess, and Nicolas Suzor. 2020. Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence* 26, 2 (2020), 237–252.
[23] Elizabeth Dwoskin, Nitasha Tiku, and Heather Kelly. 2020. Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show. https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/. Accessed by 4/10/2022.
[24] Elizabeth Dwoskin, Jeanne Whalen, and Regine Cabato. 2019. Content moderators at YouTube, Facebook and Twitter see the worst of the web — and suffer silently. https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/. Accessed by 07/13/2022.
[25] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 345–363.
[26] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.
[27] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems* 26, 1 (2009), 49–72.
[28] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.
[29] Ysabel Gerrard and Helen Thornham. 2020. Content moderation: Social media's sexist assemblages. *new media & society* 22, 7 (2020), 1266–1286.
[30] 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. Accessed by 04/18/2022.
[31] Tarleton Gillespie. 2018. *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. 1–288 pages.
[32] Robert Gorwa. 2019. What is platform governance? *Information, communication & society* 22, 6 (2019), 854–871.
[33] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
[34] Jenni Hokka. 2021. PewDiePie, racism and Youtube's neoliberalist interpretation of freedom of speech. *Convergence* 27, 1 (2021), 142–160.
[35] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.
[36] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[37] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[38] Shagun Jhaver, Seth Frey, and Amy Zhang. 2021. Designing for Multiple Centers of Power: A Taxonomy of Multi-level Governance in Online Social Platforms. *arXiv preprint arXiv:2108.12529* (2021).

[39] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

[40] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[41] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2023. A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–34.

[42] Tiffany D Joseph and Laura E Hirshfield. 2011. 'Why don't you get somebody new to do it?' Race and cultural taxation in the academy. *Ethnic and racial studies* 34, 1 (2011), 121–141.

[43] Makena Kelly. 2020. Inside Nextdoor's 'Karin Problem'. https://www.theverge.com/21283993/nextdoor-app-racism-community-moderation-guidance-protests. Accessed by 04/11/2022.

[44] Christopher Kelty. 2005. Geeks, social imaginaries, and recursive publics. *Cultural Anthropology* 20, 2 (2005), 185–214.

[45] Donald F Kettl. 2000. The transformation of governance: Globalization, devolution, and the role of government. *Public administration review* 60, 6 (2000), 488–497.

[46] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "Eternal September" How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1152–1156.

[47] Charles Kiene, Kenny Shores, Eshwar Chandrasekharan, Shagun Jhaver, Jialun "Aaron" Jiang, Brianna Dym, Joseph Seering, Sarah Gilbert, Kat Lo, Donghee Yvette Wohn, et al. 2019. Volunteer work: Mapping the future of moderation research. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 492–497.

[48] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* (2012), 125–178.

[49] Adam Klein. 2012. Slipping racism into the mainstream: A theory of information laundering. *Communication Theory* 22, 4 (2012), 427–448.

[50] Jan Kooiman. 2003. *Societal governance*. Springer. 229–250 pages.

[51] Yubo Kou, Xinning Gui, Yunan Chen, and Kathleen Pine. 2017. Conspiracy talk on social media: collective sensemaking during a public health crisis. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–21.

[52] Nicolle Lamerichs, Dennis Nguyen, Mari Carmen Puerta Melguizo, Radmila Radojevic, and Anna Lange-Böhmer. 2018. Elite male bodies: The circulation of alt-Right memes and the framing of politicians on Social Media. *Participations* 15, 1 (2018), 180–206.

[53] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[54] Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. 81–90.

[55] Alex Leavitt and John J Robinson. 2017. The role of information visibility in network gatekeeping: Information aggregation on Reddit during crisis events. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1246–1261.

[56] Alex Leavitt and John J Robinson. 2017. Upvote My News: The Practices of Peer Information Aggregation for Breaking News on reddit. com. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–18.

[57] Zeus Leonardo. 2013. *Race frameworks: A multidimensional theory of racism and education*. Teachers College Press.

[58] Corinne Lysandra Mason. 2016. Tinder and humanitarian hook-ups: The erotics of social media racism. *Feminist Media Studies* 16, 5 (2016), 822–837.

[59] Adrienne Massanari. 2017. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (2017), 329–346.

[60] Ariadna Matamoros-Fernández. 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society* 20, 6 (2017), 930–946.

[61] Ariadna Matamoros Fernandez. 2018. Inciting anger through Facebook reactions in Belgium: The use of emoji and related vernacular expressions in racist discourse. *First Monday* 23, 9 (2018), Article–number.

[62] Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television & New Media* 22, 2 (2021), 205–224.

[63] J Nathan Matias. 2016. Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 1138–1151.
[64] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.
[65] Aiden R McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Ann Cranefield. 2016. Controlling bad behavior in online communities: An examination of moderation work. (2016).
[66] Amanda Menking and Ingrid Erickson. 2015. The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 207–210.
[67] Lisa Nakamura. 2012. Don't hate the player, hate the game: The racialization of labor in World of Warcraft. In *Digital Labor*. Routledge, 195–212.
[68] Casey Newton. 2020. Leaving content moderation to volunteers is empowering racists. https://www.theverge.com/interface/2020/6/9/21283442/content-moderation-racism-facebook-reddit-nextdoor-karen. Accessed by 04/07/2022.
[69] Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.
[70] David C Oh. 2016. "Payback for Pearl Harbor" Racist Ideologies Online of Karmic Retribution for White America and Postracial Resistance. *Journal of Communication Inquiry* 40, 3 (2016), 247–266.
[71] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1114–1125.
[72] Reddit. 2020. Reddit Content Policy. https://www.redditinc.com/policies/content-policy. Accessed by 10/15/2020.
[73] Reddit. 2022. Promoting Hate Based on Identity or Vulnerability. https://www.reddithelp.com/hc/en-us/articles/360045715951. Accessed by 03/10/2022.
[74] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6174–6184.
[75] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation*. Ph.D. Dissertation. University of Illinois at Urbana-Champaign.
[76] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).
[77] Minna Ruckenstein and Linda Lisa Maria Turunen. 2020. Re-humanizing the platform: Content moderators and the logic of care. *New media & society* 22, 6 (2020), 1026–1042.
[78] Aaron Sankin. 2017. How activists of color lose battles against Facebook's moderator army. https://revealnews.org/article/how-activists-of-color-lose-battles-against-Facebooks-moderator-army/. Accessed by 11/27/2020.
[79] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and A Noah Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.
[80] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5477–5490.
[81] William Sattelberg. 2020. The Demographics Of Reddit: Who Uses The Site? https://social.techjunkie.com/demographics-reddit/. Accessed by 10/12/2020.
[82] Nathan Schneider. 2021. Admins, mods, and benevolent dictators for life: The implicit feudalism of online communities. *New Media & Society* (2021), 1461444820986553.
[83] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Z Tan, and Amy X Zhang. 2021. Modular politics: Toward a governance layer for online communities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.
[84] Joseph Seering. 2020. Reconsidering Community Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact* 3 (2020).
[85] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2020. Metaphors in moderation. *New Media & Society* (2020), 1461444820964968.
[86] 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*. 111–125.
[87] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.
[88] Bryan C Semaan, Scott P Robertson, Sara Douglas, and Misa Maruyama. 2014. Social media supporting political deliberation across multiple public spheres: towards depolarization. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1409–1421.

[89] Aaron Shaw and Benjamin M Hill. 2014. Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication* 64, 2 (2014), 215–238.
[90] Eugenia Siapera. 2021. AI Content Moderation, Racism and (de) Coloniality. *International Journal of Bullying Prevention* (2021), 1–11.
[91] Eugenia Siapera and Paloma Viejo-Otero. 2021. Governing hate: Facebook and digital racism. *Television & New Media* 22, 2 (2021), 112–130.
[92] Monika Singh, Divya Bansal, and Sanjeev Sofat. 2016. Behavioral analysis and classification of spammers distributing pornographic content in social media. *Social Network Analysis and Mining* 6, 1 (2016), 41.
[93] Tim Squirrell. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society* 21, 9 (2019), 1910–1927.
[94] Sam Srauy and John Cheney-Lippold. 2019. Realism in FIFA? How social realism enabled platformed racism in a video game. *First Monday* (2019).
[95] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology: An overview. (1994).
[96] Lucy Suchman. 2011. Anthropological relocations and the limits of design. *Annual Review of Anthropology* 40 (2011), 1–18.
[97] Shannon Sullivan. 2014. *Good white people: The problem with middle-class white anti-racism.* Suny Press.
[98] Tiziana Terranova. 2000. Free labor: Producing culture for the digital economy. *Social text* 18, 2 (2000), 33–58.
[99] Nanna Thylstrup and Zeerak Waseem. 2020. Detecting 'dirt'and 'toxicity': Rethinking content moderation as pollution behaviour. *Available at SSRN 3709719* (2020).
[100] Eve Tuck and K Wayne Yang. 2014. R-words: Refusing research. *Humanizing research: Decolonizing qualitative inquiry with youth and communities* 223 (2014), 248.
[101] u/des1g_. 2019. Updated list of most popular (US) city subreddits, ranked. https://www.reddit.com/r/ListOfSubreddits/comments/drrp9g/updated_list_of_most_popular_us_city_subreddits/. Accessed by 7/21/2021.
[102] u/Osiris32. 2016. An updated list of the most popular city subreddits. https://www.reddit.com/r/redditlists/comments/3yusqs/an_updated_list_of_the_most_popular_city/. Accessed by 11/17/2020.
[103] Galen Weld, Amy X Zhang, and Tim Althoff. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1121–1132.
[104] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
[105] Qunfang Wu, Louisa Kayah Williams, Ellen Simpson, and Bryan Semaan. 2022. Conversations About Crime: Re-Enforcing and Fighting Against Platformed Racism on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–38.
[106] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: building governance in online communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 365–378.
[107] Haiyi Zhu, Robert E Kraut, and Aniket Kittur. 2014. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 281–290.

Received July 2022; revised January 2023; accepted March 2023