Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?

Farhana Shahid
Cornell University
Ithaca, United States
fs468@cornell.edu

Aditya Vashistha
Cornell University
Ithaca, United States
adityav@cornell.edu

ABSTRACT

Social media platforms use content moderation to reduce and remove problematic content. However, much of the discourse on the benefits and pitfalls of moderation has so far focused on users in the West. Little is known about how users in the Global South interact with the humans and algorithms behind opaque moderation systems. To fill this gap, we conducted interviews with 19 Bangladeshi social media users who received restrictions for violating community standards on Facebook. We found that the users perceived the underlying human-AI infrastructure to imbibe coloniality in the form of amplifying power relations, centering Western norms, and perpetuating historical injustices and erasure of minoritized expressions. Based on the findings, we establish that the current moderation systems propagate historical power relations and patterns of oppression, and discuss ways to rethink moderation in a fundamentally decolonial way.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

decoloniality, moderation, care, Global South

ACM Reference Format:
Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581538

1 INTRODUCTION

Social media platforms have witnessed an unprecedented growth in users based in the Global South, who are highly vulnerable to harmful content like misinformation and hate speech [131]. Although the platforms use content moderation to reduce and remove problematic content, the monolithic moderation systems often fail to account for large sociocultural differences between users in the Global South and users in the West. Since most large social media platforms are based in the United States, their moderation policies are heavily influenced by the Western norms, particularly the First Amendment of *free speech* [52]. Although the platforms have standardized their moderation policies globally in response to the growing number of users in the Global South, these users are still controlled by the Western hegemony which decides what forms of expressions are *acceptable* locally [52]. Even when the platforms attempt to make the moderation more attuned to the needs of the users in the Global South, they often do so by recruiting local moderators from the existing power hierarchies and government agencies [60], perpetuating a vicious cycle of crackdown on political dissidents in the name of moderating rumors, misinformation, and maintaining law and order [132, 150].

A growing evidence points to not only apathy towards users in the Global South but also blatant discrimination against them. For instance, Lyons [87] points that Facebook endangered users in the Global South by using them as *test subjects* to assess their underdeveloped moderation policies before applying them to handle the chaos of the US election. Similarly, many large social media platforms have neither any moderation tool nor oversight mechanism to prevent problematic content in popular non-Western languages [122]. For example, Twitter's new Bodyguard tool, which protects users from cyber bullying, hate speech, and toxic content, is only available in English, French, Italian, Spanish, and Portuguese [115]. Such disparities in moderation have led to tragic consequences and grave abuses in the Global South [4, 118].

While a growing scholarship has started examining the inequities and biases in content moderation, to date, most research is heavily skewed towards and informed by the experiences of users in the West [64, 110, 147, 151]. Little is known about how users in the Global South engage with and experience the opaque content moderation systems. To fill this gap, we sought to answer the following research questions:

- **RQ1:** How might content moderation systems propagate historical power relations and patterns of oppression?
- **RQ2:** What steps can we take towards a *fundamentally decolonial* content moderation system?

To answer these questions, we conducted semi-structured interviews with 19 Bangladeshi Facebook users, who directly engaged with the underlying content moderation processes for violating community guidelines. Our analysis found several anomalies in current moderation practices that distressed Bangladeshi Facebook

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581538

users and shaped their online expressions. Our participants perceived that Facebook's content moderation system imbibes coloniality by amplifying power relations, centering Western norms, perpetuating historical injustices, and censoring minoritized expressions. Contrary to the Western users, who took their right to freedom of speech for granted against content removal [147, 151], our participants felt that they were constantly subjected to scrutiny by Facebook based on the Western social norms and perceived Facebook's moderation policies to be far removed from the local sociocultural norms and values. They expressed concerns about how their diverse cultural and linguistic expressions were misunderstood by the underlying algorithms and perceived them to be trained on the Western languages by the Western researchers, who might lack appreciation and understanding of users' local context. Even when unfair restrictions were imposed on users, largely because the current policies failed to account for local norms and sensitivities, Facebook rarely explained to users why they were restricted and gave limited avenues to appeal the decision. Thus, users were left on their own to make sense of the restrictions and had to adopt algospeak and anti-programming strategies to counteract repressive moderation policies.

Drawing on these findings, we use *decoloniality* as a lens to show how the Western social media platforms like Facebook function as the modern *metropoles* of *digital colonialism*, peripheralize the needs of the users from the Global South communities like Bangladesh, and exclusively control the code of appropriate online conduct. Our participants felt that even though platforms benefit economically from the expansion of their market in the Global South countries like Bangladesh, platforms do little to serve users like them fairly and equally during the entire content moderation pipeline. Moreover, platforms exploit the ill-regulated, cheap workforce in the Global South to outsource moderation responsibilities without adequately upgrading the working condition of the local moderators [135] and investing in new technologies to moderate local content. We provide design recommendations towards a fundamentally decolonial content moderation pipeline that might offer a level playing field and equitable access to moderation related information, support, and services for the users, whose concerns and needs are often ignored in the moderation infrastructure. To do this, we turn to the value of *care* in HCI design [19, 58, 76, 144] and propose Bellacasa's three dimensional notion of care [41] as an ethical-political commitment to advocate for designing decolonial content moderation systems. The key contributions of our work are as follows:

- A qualitative study that uses *decoloniality* as a lens to provide a critical and nuanced understanding of how the Western hegemony in current content moderation infrastructure perpetuates digital colonialism and enables systematic oppression and discriminatory experience for Bangladeshi Facebook users.
- Reflections grounded within *care* to reimagine a fundamentally decolonial content moderation infrastructure.

In summary, we situate our work as a crucial first step towards examining the colonial elements in current content moderation infrastructure that deeply impacts the experiences of non-Western users. As one size doesn't fit all, the experiences and perceptions of Bangladeshi Facebook users might not equally apply to the diverse Global South regions. However, we expect a decolonial lens will steer conversation in the right directions to probe into the experiences of users from the diverse Global South communities and identify their unique needs and circumstances that need to be addressed and cared for to ensure a fair and just moderation system.

2 RELATED WORK

Content moderation has become a hotbed for debate among the researchers, policy makers, journalists, human rights organizations, and users. Although platforms have set up elaborate processes to moderate content, the opacity around the moderation pipeline and its outcomes [82] are often at odds with platforms' proclaimed intent and users' expectations. We situate our work first by narrating the moderation pipeline used by different social media platforms. We then discuss past work on how users perceive and experience the effects of content moderation. Finally, we use decoloniality as a lens to see how the effects of colonialism are broadly understood within HCI research and particularly in the case of content moderation.

2.1 Content Moderation Pipeline

Given the massive scale of content generated on social media, most platforms rely on automated models to identify and remove harmful content before it reaches users [24, 66]. When additional inputs are required, the platforms escalate the automatically flagged content to human moderators for further review. These moderators are either full-time employees of the platforms, external content reviewers employed at third party organizations, or people working voluntarily [8, 24]. Once the platforms identify a post to be problematic, they either reduce its distribution or remove it, and take punitive measures against the offending user [67, 92, 119]. For example, Facebook and YouTube follow a *strike* system to count user violations and usually give a warning to users on their first violation [67, 92]. For repeat offense added restrictions are imposed on users which may eventually lead to a permanent ban from the platform. While some platforms (e.g., Reddit) enact these penalties without notifying users [72], others (e.g., Facebook, Twitter) use notifications or emails to inform users of violations [28, 29]. Some platforms (e.g., Facebook, Reddit, Twitter) allow content violators to appeal the restriction if they disagree with the platform's decision, though there is a variance across platforms in what kinds of restrictions can be appealed [9, 29]. However, the appeal process largely remains opaque across all platforms.

2.2 Users' Perceptions and Experiences of Content Moderation

Content moderation encapsulates an invisible infrastructure of human labor and computing advances, underlying parts of which (e.g., content removal, restrictions, appeal) are often only visible to norm violators [56, 139]. Prior work shows that users support moderation to safeguard others from harmful content unless they fear an attack on individual freedom [120] or more scrutiny than their political counterparts [64]. However, the opaqueness of the

moderation processes often make them seem biased and inconsistent [55, 64, 72, 147, 151]. Several factors—such as moderation equity, consistent implementation, and visibility of algorithmic decision making—shape users' notion of fairness [88]. Users perceive moderation decision to be legitimate when it aligns with their preferences, is made by expert panels, and properly communicated to them [70, 110]. For example, letting users appeal without explaining *why* their content is removed negatively impacts users' perceptions of fairness, accountability, and trustworthiness of algorithmic decisions [147]. Individual differences also shape such perceptions, given that some users tend to trust automated moderation more than human moderators [98]. Even the nature of moderation, e.g., community-based moderation, makes platforms appear more transparent to users than the commercially moderated platforms [32].

Recently, several researchers have studied the experiences of historically marginalized users with content moderation and found the underlying moderation processes to be ableist, sexist, racist, and homophobic. Platforms tend to disproportionately moderate content from disabled, fat, queer, Black users, and women of color [12, 21, 25, 80]. For example, Gerrard and Thornham [54] examined how social media's *sexist assemblage* perpetuate predefined gender roles to police content related to women and their bodies. Moreover, prior work with users from LGBTQ [14, 64] and pro-eating disorder communities [30, 50] show that content removal and restrictions disparage marginalized users, take away valuable support resources from them, and exacerbate power disparities by privileging normative expressions [85, 121]. To address these concerns, researchers have demanded external visibility of the moderation process [141] and proposed to restructure moderation by incorporating the will of people [13], constitutional values [40], civics oriented approaches [48], and restorative justice sensibilities [125]. While these approaches are strides in the right direction, most of these frameworks are informed by the inputs and opinions of the Western social media users. Even though users in the Global South represent the majority of the user base of the Western social media platforms, little is known about their experiences and engagement with content moderation systems. As a *first* step to address this critical gap, we use a *decolonial* lens to critically examine the experiences and impacts of content moderation on users in Bangladesh who violated Facebook's community standards. We now present scholarly work that dissect different dimensions of coloniality within computing and HCI, and discuss how they tie back to online content moderation.

2.3 Coloniality, Computing, and Care

The history and legacy of marginalization and colonization are deeply intertwined. The place of our study, Bangladesh—part of the *Bengal* province in pre-independent India—was first under the imperialism of British East India Company and later British colonial rule for almost two centuries [31]. Critical scholar Ashis Nandy [101] scrutinized how the imperial hegemony of the British produced social hierarchies that enabled the West to project its political dominance over the cultural fabrics of the Indian subcontinent as they tried to *civilize* the locals. Even after attaining freedom from the colonial rulers, the power dynamics between the advantaged and disadvantaged in socio-economic and political spheres still persist for the historical dispossession, appropriation, and extraction of knowledge, labor, and resources from the colonies [20]. According to Quijano [116, 117], this power dynamics is a product of colonial exertion and control over the sociopolitical and economic structures of the colonized along the dimensions of authority, gender and sexuality, and knowledge and subjectivity.

The colonial structures of power, control, and hegemony also shape the design of technological systems–either as an all encompassing, *universal* solution to problems that ignore their local nuances or with the pre-held notion that people from the former colonies lack the skills to solve their own problems [138]. In the face of an expansionist colonial impulse to incessantly computerize the modern world, Mignolo stresses to affirm the modes and practices that have been historically denied by the dominance of Western norms [95]. This has probed many HCI scholars to use *decoloniality* as a lens to critique Eurocentric imposition of a singular form of knowledge as *universal* and superior to others [5, 11, 43, 61]. For our study, we used a decolonial lens as it recognizes the fact that coloniality still survives under the cloak of *modernity* and perpetuates its harms by colonial control of power, knowledge, and being [103, 116]. For instance, Bidwell [22] criticized how the design of many technological interventions for low-resource communities in developing regions embody coloniality by imbibing the Western logic of individuality while ignoring the sociality of the Global South. Despite the benefits that the design and production of technologies in the West are cashing out from the resources and labor of the former colonies, these technologies are complicit in erasing and dominating the sociocultural fabric of the colonies [15]. This colonial appropriation is reproduced on modern digital platforms by commodifying and capturing human relations in the forms of data [35]. For example, data colonialism [34] and the promise of algorithmic utopia [6] have accumulated unrestrained power in the hands of predictive AI systems that perpetuate biases and harms disproportionately against the historically marginalized populations [33, 106, 113, 133].

The coloniality of power also shapes the content moderation systems used by the Western social media platforms to govern the forms of speech that is *acceptable* in an increasingly online world. Building on María Lugones' [86] call to revisit coloniality through the lens of race, gender, and sexuality, Siapera [129] used decoloniality as a lens to argue that the content moderation processes fail to identify and address racism as a structure of colonial power, take little input from racialized users in deciding policies against racist hate speech, and exploit unpaid labor (i.e., user reported hate speech) to train AI models. We contribute to this nascent line of work by unpacking the colonial elements in moderation pipeline that broadly shape the experiences of the Bangladeshi Facebook users. Through our analysis, we find that Facebook's community standards disregard the sociocultural norms of Bangladeshi users and are used to police and censor user activities from abroad. Decolonial lens helps us unravel how modern social media platforms like Facebook continue to exploit users from the Global South countries like Bangladesh by controlling the rules of online expression that do not take into account local norms and values.

To address such colonial aggression, many HCI and critical computing scholars have advocated for shifting the center of power

and knowledge to the peripheries [22] and embracing cultural pluralism into design [47, 93, 127, 152]. For this to be accomplished, decolonial processes presuppose care and commitment in designing technologies for, *and with*, the underserved communities [37]. Many HCI scholars have adopted *care* as a framework to develop a rich understanding of the dynamics in digital spaces and collaborative work [18, 143–145], examine sociotechnical obligations to normative and universal moral principles [105], unravel the politics of invisibility [16, 99], and design equitable and responsible sociotechnical systems [19, 91]. Scholars like Yu et al. [154] have described content moderation as a form of *care work* done by the moderators to *maintain* a safe online space by shouldering the responsibility to remove harmful content. However, critical scholar Puig de la Bellacasa [41] has argued looking beyond care as a maintenance work and insists to navigate the tensions and relations along the three-dimensions of: (1) ethics/politics, (2) labor/work, and (3) affect/affections. Bellacasa has discussed that the ethical and political values of care posit the question *"how to care"* and inquire into different types of labor that make the care work possible. She dissects the *ethical obligation* ingrained within everyday labor of maintenance in technological spaces that enables the *becoming* of such digital spaces. She also nudges to think about the disempowering effects of such obligatory care work that usually falls upon the less privileged and how its outcomes affect them the most.

Even though the platforms proffer moderation as a panacea for safe online space, our findings reveal the underlying politics behind how platforms enact their notion of *care* by imposing a universal ethics on all users. This enables persistent discrimination in the division of labor that makes moderation possible, leads to disproportionate outcomes of moderation for different user groups, and causes affliction to users with unaccommodating restrictions and appeal processes. Therefore, it is necessary that we study content moderation holistically, by unraveling its underlying politics, disentangling the labor that goes behind it, and assessing how it affects users' online being. Our work contributes to the scholarship on care by integrating Bellacasa's framework as an alternative design lens to improve the fundamental relational quality of content moderation. Using the multi-dimensional notion of care, we discuss how to rethink moderation in a fundamentally decolonial way to facilitate respectful recognition of diverse knowledge, values, and worldviews.

3 METHODS

We conducted semi-structured interviews with 19 Bangladeshi social media users who received restrictions from Facebook for violating community standards. Participation in our study was voluntary and the study protocol was approved by the Institutional Review Board at the authors' institutions.

Participant Recruitment. We used a combination of snowball and convenience sampling to recruit participants. We advertised our study in different Bangladeshi Facebook groups to recruit users with first-hand experience of community standard violations. Some participants also shared the news within their Facebook circles and helped us recruit more participants. We continued recruiting participants until the responses reached theoretical saturation [111]. We contacted people who were interested in participating in our research via Messenger, shared details of our study with them, and scheduled their interviews. Due to the ongoing COVID-19 pandemic, we conducted the interviews online via video conferencing platforms of our participants' choice.

Semi-Structured Interviews. We designed our interview protocol to learn about participants' experiences with content moderation, community standard violations, and subsequent restrictions and appeal process. All correspondence and the interviews with the participants were conducted in Bengali, their native language. First, we shared the informed consent script with our participants and requested their consent. After users agreed, we asked them details about the content that was flagged by Facebook, the restrictions imposed on them, and the process that followed. While some users shared screenshots of their records of past violations, others either did not know how to access the restriction history or could not locate the record as it was made unavailable by Facebook after a certain period. We then asked users how they felt about the restrictions imposed upon them, the difficulties they faced while trying to appeal, and the shortcomings in current moderation practices. Finally, we asked users about their expectations from social media platforms, including changes they would like to see in current moderation policies. After each interview, we revised our questions to add new probes, stopping when participants' responses reached saturation. Each interview lasted approximately 50 minutes, and was audio-recorded with the consent of the participants.

Data Collection and Analysis. We collected around 17 hours of interview data and 38 screenshots of users' violation history. We transcribed and translated the interviews into English and coded them using inductive thematic analysis [62]. We took multiple passes on the transcribed data and users' records of past violations to conduct open coding. We avoided using any pre-supposed codes and instead let the codes emerge freely from the data. During the analysis, all authors met regularly to discuss the emerging codes, develop preliminary codebook, review and update the codes, resolve the disagreements through peer debriefing [36], and develop the categories and themes. The prolonged engagement with the data helped us establish credibility. After multiple iterations through data, our collaborative analysis produced 46 codes. We further clustered the codes into three high level themes around user perceptions of moderation, subverting the harms of moderation, and rethinking moderation.

Participant Demographics. Our sample had 19 participants, all of whom identified as male except one (see Table 1). On average, our participants were 32 years old (SD: 7 years), with ages ranging between 22–50 years. A majority of them (N=13) were from the capital city Dhaka and the rest were Bangladeshi citizens living in the United States, United Kingdom, Canada, and Germany at the time of the interview. All participants were highly educated; ten with a bachelor's degree, seven with a master's degree, and two with a doctoral degree. Our sample had people from diverse professions, including software engineers, students, journalists, teachers, and government employees, among others. All participants owned a smartphone, including at least another electronic device, such as a laptop, desktop, or iPad. All of them regularly used Facebook, Messenger, and other social media platforms like Instagram, WhatsApp, Twitter, and LinkedIn, among others.

Table 1: Pseudonym, reported age, gender, and profession of all the participants.

Name	Age	Gender	Profession	Name	Age	Gender	Profession
Mithila	27	Female	Software Engineer	Omar	38	Male	Businessman
Faruk	37	Male	Journalist	Himel	29	Male	Student
Shafiq	39	Male	Journalist	Zahid	27	Male	Researcher
Rashed	30	Male	Corporate Manager	Ehsan	41	Male	Research Scientist
Tanjib	22	Male	Student	Hamid	22	Male	Student
Minhaj	36	Male	Teacher	Zubayer	50	Male	Patent Examiner
Anidnya	29	Male	Software Engineer	Kabir	39	Male	Civil Engineer
Asad	29	Male	Student	Saif	27	Male	Software Engineer
Iftekhar	42	Male	Software architect	Rafat	22	Male	Student
Razzaque	30	Male	University Teacher				

Positionality. We embrace Smith's [134] proposed self-reflexive guideline to approach decolonial research studies. All authors are from historically colonized regions in the Global South, have first-hand experience of being in spaces shaped by coloniality, and have extensive experiences of conducting research in the Global South. The first author, who conducted the interviews is a native Bengali speaker. Her urbanity, sociocultural, and educational backgrounds are on par with the participants. Even though we do not have personal experiences of going through restrictions resulting from content violations, our prior engagement and shared backgrounds with the community helped us develop a nuanced understanding of the underlying sociocultural constructs that often make users susceptible to content violation and shape their subsequent reactions. Even though we come from historically colonized regions, we are affiliated with institutions that's built on and with the money obtained from the forcefully appropriated lands of the Indigenous people. We relate to Villenas's [149] *"feet in both world"* for belonging to historically colonized communities and the institutions that were benefited by colonial agenda. We posit this work as part of a broader decolonizing agenda within HCI research [7, 84] and affirm the centrality of land reparations in decolonization process [146].

4 FINDINGS

We first present how current content moderation systems sustain colonial domination of power and disparage everyday expressions of users in Bangladesh that do not conform to the Western standards (§4.1). We then discuss ways in which participants coped with the harms of moderation (§4.2). Finally, we describe participants' views on implementing content moderation in a way that is more attuned to their needs and aspirations (§4.3).

4.1 Users' Experiences with Moderation: An Extension of Colonialism

All participants felt harassed by Facebook's insensitive content moderation decisions irrespective of whether they lived in Bangladesh or abroad. They perceived the restrictions for violating community standards to be severe, punitive, unfair, and unjust. Some participants received a warning after being flagged for violation, while others had their content removed, profiles/pages deleted, and restrictions placed on what they can (and cannot do) on Facebook. We now unpack how participants struggled with content moderation and why they viewed it as a modern tool to enforce the Western power hegemony.

4.1.1 ***Otherization by the Western Community Standard.*** Participants felt that Facebook prioritizes the Western liberal values on issues, such as feminism, atheism, privacy, and gender conformity, often discounting distinct social structures and dynamics that dictate local values and sentiments. They asserted that what is acceptable in one society might be considered inappropriate in another and gave several examples when moderation, fueled by the Western standards, subjected users to restrictions for posting content that was in line with the local sociocultural norms. Razzaque shared:

> *"While I was visiting my village, I photographed small kids swimming into the river. They usually swim without clothes to avoid being scolded by parents. But when I uploaded these pictures, Facebook removed them saying they depict child nudity. Village kids learn to swim very early. This [swimming naked] is their natural way of being and starkly different from the Western understanding of child nudity and pornography. Removal of my post shows that the platform neither understood my culture nor respected it."*

In addition, participants shared several examples when content moderation failed to account for *"basic common sense."* For example, a user joined a Facebook group to connect with college alumni and introduced himself by mentioning that he studied Science. His comment was removed for violating community standards on cybersecurity (see Figure 1(A)). He had no clue why the comment violated the standards, especially since Facebook shared no explanation. He assumed that he missed a space between *'3.Science'* which might have been misinterpreted to be an unsecured website. Another participant Ehsan described how a Facebook group of 90,000 members, created for providing support during health emergencies, was deleted after several posts on blood donation requests were flagged for breaching user privacy when group members shared their contact details to coordinate with each other.

> *"I was a moderator of the group. One morning we found that the group was gone. The admin received an email saying several posts breached user privacy. Since our group is about connecting people, many users posted blood donation requests and also provided their contact number and hospital addresses, which was seen as a breach of personal information. However, if they*

don't share contact numbers, how will blood donors reach out to them?"

This example shows how the moderation process enforced the Western notion of privacy on Bangladeshi users in a context where it is socially and culturally acceptable to share one's personal information publicly. Participants felt that *universalizing* such highly contextual norms allows only one acceptable way to speak and polices user expressions that fundamentally differ from the Western norms, but are important in the historical and cultural context of Bangladesh. In another example, Shafiq expressed his frustration when the pictures depicting war crimes against Bangladeshis were removed for violating community standards:

"A page called [anonymized] posted some historic images of the liberation war showing the brutality Pakistani army inflicted upon the Bengalis [Bangladeshis]. Facebook removed the images saying that they depict violence. This way they are trying to censor our country's history and hush the war crime of genocide."

Our participants were not only *"othered"* when Facebook failed to acknowledge their historical injustices and sufferings, but felt unsafe in discussing mundane, everyday things without being scrutinized. For example, Hamid's friend posted in a Facebook group of fellow students expressing relief at the end of final exams, *"Who is willing to burn effigies of the semester?"* When Hamid commented, *"Will burn tomorrow"*, the comment was removed for inciting violence. In another incident, Zahid's friend posted, *"Looking for status updates to attract beautiful girls but not finding any. Feeling frustrated."* When Zahid mocked his friend commenting, *"Who will be attracted to a fat pig like you?"* (see Figure 1(B)), his comment was removed for bullying and harassment and his Facebook activities were restricted for three days. He explained:

"We always roast each other this way. But now Facebook is restricting our interaction with friends, especially when my friend is absolutely fine with my comments. We want to interact with friends the same way we do in real life."

These examples show that the current moderation systems not only discount the sociocultural norms of the local community but also make little effort to understand the surrounding context and relationship dynamics. The *"lack of common sense"* in moderation led many participants to believe that the moderation pipeline is fully automated, a sentiment that has been echoed by social media users in the West as well [147]. Our participants emphasized how it would be impossible for human moderators to not *"understand the sarcasm or humor"* in the flagged posts. In addition to missing or misinterpreting the context of the posts, our participants emphasized the linguistic discrimination enforced by the moderation algorithms. They complained that the algorithms were putting restrictions on users for using curse words in Bengali, even though the posts containing equivalent English expletives were allowed. For example, Faruk posted a Facebook status narrating his interaction with a taxi driver where the driver called his own son *"kuttar baccha"* (son of a bitch) for failing the exams. Faruk's post, where he quoted the driver, was removed for violating community standards and his Facebook activities were restricted for seven days. He was surprised that despite using quotes, the algorithm failed to understand that he was narrating a conversation instead of using profanity against others.

Participants expressed that even though Facebook makes money by selling user data globally, it only invests in developing AI models for the English language while ignoring the needs of non-English speaking users. Participants with a technology background pointed that the impressive availability of automated moderation tools for the Western languages and the domination of the Western researchers in designing moderation technologies reaffirm colonial *"control of power."* They expressed that Facebook should stop judging users' posts on local issues based on the standards and technologies primarily developed for and by people in the West. For example, Shafiq shared how Facebook's moderation policy around racial hatred is mostly informed by the racial issues in the West, which do not translate well in Bangladesh where the dominant Muslim majority is not divided into any race. He elaborated:

"You and I became educated recently and learnt that we should not make comments based on skin color. But my grandmother can comment affectionately, 'Hey your son [who has a dark skin color] is a black diamond [as an appreciation for child]. She doesn't understand racism. How can you moderate her affection using the Western community standards?"

Participants were also confused how content moderation systems handled religious content. For example, Asad shared a picture of the Quran (the holy book of the Muslims) placed in the lap of a Hindu goddess with the caption: *"No religion teaches to disrespect the holy book of another religion"* (see Figure 1(C)) which was later removed. Since Facebook did not explain why the post violated community standards and how it could instigate hatred against the Hindus, Asad considered the action as *"Islamophobic."* Like him, several participants were confused by inconsistent removal of hate speech and lack of communication to users about how Facebook defines *harmful* content. Secular participants complained that Facebook often prohibits constructive criticism of Islam in the name of Islamophobia, but allows many pro-Islamist, divisive posts. Iftekhar shared:

"Islamic militants openly promote communal violence against the Hindus and threaten to slaughter and murder us, the activists. When we report these activities, Facebook says the content doesn't violate community standards. I suspect they are trying to push countries like ours towards Islamic militancy while silencing liberal voices."

These examples show why many participants perceived moderation as dictatorial, silencing their voices. They pointed that the community standards are usually decided in silos by people sitting at the top of the power hierarchy in *"Silicon Valley companies without engaging the voices from the ground."* They questioned why Facebook gets to decide which content is representative of *their* community, what is the appropriate way to use *their* language, and how they should interact with *their* friends. They felt that the platforms like Facebook exploit moderation to advance their colonial agenda of free market policy to control the communities globally. They felt that Facebook is driven to maximize engagement and as a result, purposely neglect negative comments on celebrity profiles to fuel conversations. Similar belief also prevailed among the US

youth, who doubted platforms' integrity in maintaining a fair resolution and believed that social media corporations are motivated by profit [126]. Razzaque elaborated:

> *"Facebook simply asks 'What's on your mind?' Like colonizers they just want to profit by taking away from us without giving anything in return. This attitude needs to be changed."*

Some participants even compared Facebook to public spheres like tea stalls where people gather to chat, shop, and be entertained. Similar to such offline settings, they expressed there shouldn't be any global standards for public discourse on social media. Shafiq elaborated:

> *"How would you set a standard for the world? If we want to establish a global community standard, are we saying that we want to move towards one world one order? Is it the old socialist utopia for equity?"*

4.1.2 ***Inaccessible Standards and Oppressive Moderation.*** Many participants were unaware of community standards until they were restricted for violating them. They complained that Facebook never provided any resources on community standards, even when they were notified of violations. Some participants read community standards after being restricted, and found them to be vague, confusing, and lengthy. Similar to prior findings [151], participants found the standards to be written in *legalese* and expressed that simply making users read the standards won't make them accept the moderation as fair [70]. Given the standards were written in English, several participants were concerned that the standards are incomprehensible to a large population in Bangladesh, especially for the low-literate and non-English speaking users. Rashed elaborated:

> *"All users do not know English well. We may understand the standards little if they are written in English. Rickshaw pullers [drivers of local passenger cart] also use Facebook and they won't be able to read the English guideline."*

The lack of effort to make community standards accessible led many participants to doubt Facebook's integrity and willingness to keep the community safe. Participants often gave example of Facebook's negligence in preventing misinformation during the Rohingya genocide and believed that the verbose and legalese community standards are designed to shift the blame elsewhere. Many participants criticized Facebook's lack of transparency in moderating content. For example, Facebook often deleted user content, accounts, pages, or groups for content violation, but did not provide any reference to the content that violated the standard (see Figure 1(D)). Not only participants found moderation to be arbitrary, but also deeply offensive and frustrating. Zubayer shared:

> *"They notified me saying my post violated community standards but didn't show which post. Then I checked my Activity Log and found the mention of 'this content violated community standards'. But when I clicked they didn't show me the content. However, from the date and time, I could guess the post. Then I checked my timeline and found that the post was gone. I had to figure out this way."*

For community standards violation in groups, Facebook currently notifies the admins only, instead of the group member whose content is removed for violation. This lack of communication caused confusion within Facebook groups. Group admins found this to be a divisive policy. They had to face heat from the group members, who often suspected that the admins intentionally deleted their post, without knowing that Facebook removed the content. Zubayer shared:

> *"When Facebook doesn't inform the group members of their community standard violations, they assume that admins intentionally deleted their posts. This often sours the relationship among the admins and group members. Then we have to search the log to figure out if the user's post was removed by Facebook or another group admin."*

In the absence of explanations on how users' content violated community standards, participants perceived the moderation process to be a black box, similar to the social media users in the West [151], and made assumptions about why they are targeted. Participants, who were engaged in political activism assumed they were reported by their adversaries. Iftekhar noted:

> *"During the road safety movement in Bangladesh I criticized the government condemning the arrest of the journalist and activist Shahidul Alam. I did not use any swear words in my writing but still received penalties. I think, I am in the watch-list of DGFI [Directorate General of Forces Intelligence]. Whenever something bad happens in the country and I criticize the government, they always try to restrict me."*

Many participants felt that the reporting feature, which allows users to flag problematic content, is weaponized by political parties and trolls to silence users with different political and religious ideologies. Participants were shocked at how easily Facebook could be fooled with fake allegations of copyright infringement and impersonation. For example, Kabir shared that his Facebook account was suddenly deactivated after several of his personal pictures were removed for copyright violations. Later, he found that someone created a fake Wordpress website using his name where his personal pictures were posted with fake past dates than that of his original posts. He suspected that someone presented false proofs to silence him by reporting his account for impersonation and copyright violation. Several participants also complained the misuse of the reporting feature to harass high-profile Bangladeshi activists. They gave the example of journalist Tasnim Khalil and writer Taslima Nasreen, who were misreported to be dead by trolls and their Facebook accounts were memorialized by the platform [10].

Participants expressed grave concerns over Facebook's arbitrary determination of hate speech and abuse. For example, Himel shared that a Facebook user called him *'Malaun'* (a slang for the Bengali Hindus). When he replied to the comment saying, *"Did anyone stop you from sharing your opinion? What kind of behavior is this to curse someone as Malaun?"* (see Figure 1(E)), his reply was removed as hate speech and his Facebook usage was restricted for a day. In another instance of unfair moderation, Zubayer recounted that he was restricted by Facebook for criticizing their moderation policies. He posted a screenshot of his reporting a post containing misinformation about a renowned freedom fighter and wrote, *"A person called Jahanara Imam's son, the brave freedom fighter as a terrorist. But the post prevailed even after reporting"* (see Figure 1(F)). His post was removed for hate speech and he got restricted for seven days.

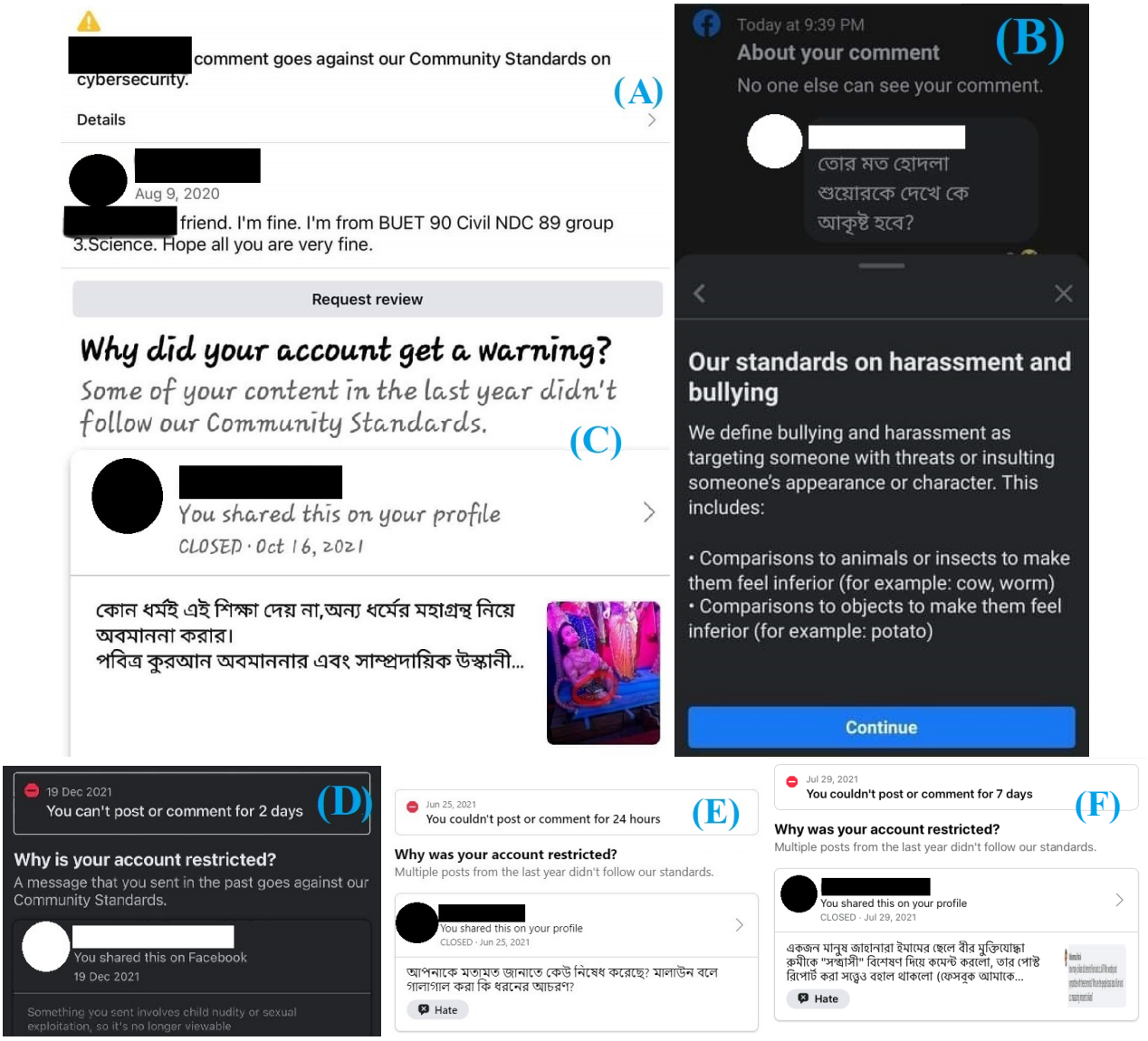

Figure 1: (A) User's self-introduction got flagged for violating cybersecurity. (B) User's comment (*'Who will be attracted to a fat pig like you?'*) on a friend's post was removed for bullying and harassment. (C) Warning given to a user who criticized a picture for showing disrespect towards Muslim's holy book *The Quran* as it was placed on the lap of a Hindu Goddess. (D) User's post that was removed for child nudity and sexual exploitation wasn't revealed to him. (E) User got restricted for calling out another user, who cursed him with a pejorative term for Bengali Hindus. (F) Facebook restricted a user for complaining against their moderation policy.

A few participants felt discriminated by moderators and group admins, who shared different political and religious ideologies than them. They suspected that the moderators are often conservatives with anti-liberation agendas. For example, Himel reported a pro-Islamist Bengali poem inciting hatred and violence against the LGBTQ community, but found it infuriating when Facebook did not remove it. Similarly, some participants complained that group admins often misused their moderation power to harass the religious minorities. Ehsan elaborated:

> *"There is a Facebook group of Bangladeshi Canadians where the admins harass other users. In a post about Islam permitting Muslim men to have up to four wives, a Hindu user commented with some reference. The admins belittled him for commenting on Muslim marriage policy despite being a Hindu and removed him from the group."*

These finding show that the lack of accountability and transparency in content moderation created a ground for exploitation,

threatened civic engagement, reduced tolerance for different viewpoints, and diminished chances for respectful online discourse.

4.2 Coping With the Invisible Harms of Moderation

We now unpack the harms that participants experienced as a result of oppressive moderation processes. We also discuss how our participants tried to cope with the injustices and harassment resulting from content moderation.

4.2.1 ***Perpetuating Harms of Moderation.*** The consequences of unfair and unjust moderation often lasted beyond the duration of the restrictions inflicted upon the users. Many participants complained that Facebook used records of past violations to restrict users from advertising on the platform and it greatly affected those who did online business. Shafiq shared:

> *"Last year during Durga Puja [a religious festival of the Bengali Hindus] I wrote against communal violence towards the Hindus but my post got reported and I was restricted for a month. This harmed me a lot as I could not advertise or sell anything from my business page. I lost money and it was severe for my family."*

Anindya, a Facebook page admin, shared that when posts on liberal topics, such as supporting feminism and condemning communal violence, got reported by conservative users, Facebook imposed restrictions on the personal accounts of all the page admnins and significantly decreased the reach of the page. Often the restrictions led to a cascading series of bans that went beyond the original platform on which the content violation happened. For example, Rafat shared that his Facebook account was disabled (Figure 2(A)) after he shared intimate content with his girlfriend over Messenger. Following this, his WhatsApp account associated with the phone number that he used for two-factor authentication on Facebook (see Figure 2(B)) and his Instagram profile that he used as a professional photographer to display his work were also disabled (see Figure 2(C, D)).

Like users in the West [151], many participants complained that they could not download their data once their Facebook page or account was disabled for violating community guidelines and thus, they lost many valuable personal data and memories. Such restrictions deeply impacted users' mental health and well-being. For example, Mithila shared that her friend, who is a suicide and abuse survivor, was restricted from using Messenger after she posted a picture with a blade, something which she often did to vent her emotions and distract herself during the moments of vulnerability. The ban temporarily disconnected her from her online support network and made it difficult for her to reach anyone for help. Like Mithila's friend, many participants self-censored their speech to avoid restrictions in future. These examples show that Facebook offered little support to users to recuperate in the aftermath of violations.

4.2.2 ***One Sided Verdict and A Threadbare Appeal Process.*** All participants were angry and annoyed that Facebook penalized them without giving a chance to explain first. Not only users felt embarrassed, but also the public nature of these restrictions was seen as an attack on their dignity. Razzaque shared:

> *"I am a teacher and I have some social acceptability. Why did they send me a warning instantly and why would they think that I violated? They could have emailed me or texted me either on Messenger or WhatsApp as they are all linked. They could have queried, 'Did you violate this community standard? What do you think?'"*

Many participants did not receive an option to appeal and were forced to accept the restrictions Facebook imposed on them. Some participants, who received the appeal option could only choose either to accept or disagree with the allegation of violation. Only three participants received a detailed appeal prompt that asked them why Facebook's decision was wrong, why they posted the content, and the surrounding social issues, linguistic and cultural aspects. However, the participants complained that they had to fill the form in a short span of time and Facebook hardly supported them to process the situation, cope with the frustration, and present cogent arguments. Overall, the process was lengthy, inconvenient, stressful, and exploitative, given the review was not even guaranteed. Zubayer shared:

> *"When I appealed Facebook told me that currently they are short of staff due to COVID and the review is not guaranteed. They will randomly select from the appeal requests and review via third party and it might take nearly 30–60 days. I was restricted only for 7 days and a 30–60 day long review process did not make any sense. Anyway I assume my case was not selected as I never heard back from them and ended up remaining restricted for 7 days."*

Participants described how Facebook doubly exploits them, first by imposing unfair and unjust restrictions and second by *"learning about mistakes in content moderation"* from the unacknowledged labor of users when they appeal the violations with little to no support. Participants also reported several anomalies in the appeal process, including receiving the decision of appeal after going through full restriction, serving restrictions even after the allegation of the violation was proven wrong, and no option to re-appeal after they accepted Facebook's allegation of the violation (see Figure 2(E)). Hamid shared:

> *"I was baffled at the sudden occurrence of the event, went on clicking continue, continue, and did not appeal. Later when I discussed with my friends I realized it [the content violation] wasn't my mistake. Then I wanted to re-appeal but didn't get any option."*

In the face of the broken appeal process, several participants sought help from their personal contacts employed at Facebook to restore their account. For example, Kabir shared that during the road safety movement in Bangladesh, several of his posts criticizing the government got reported and his Facebook account was deactivated without any prior warning. Finding no assistance from the platform, he approached one of his acquaintances working at Facebook and provided his account details. Following the acquaintance's mediation, Kabir got his account back almost a week later. Several participants compared the appeal process to the bureaucratic systems of the colonial era which were beyond the reach of the ordinary masses in Bangladesh. Participants felt being trapped by the whims of the humans and algorithms behind content moderation systems. Moreover, they duly noted that not everyone has

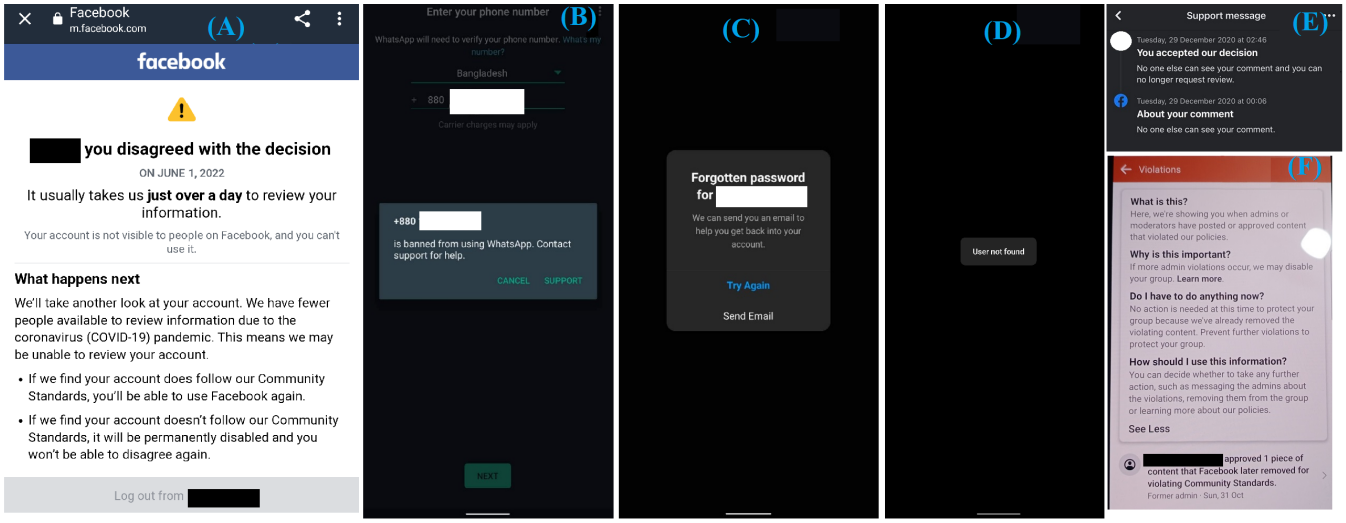

Figure 2: (A) User's Facebook account is disabled. (B) User's WhatsApp account opened with the phone number that is linked to the disabled Facebook profile in (A) got banned. (C) Forgotten password prompt shown to the user when he tried to access the Instagram account linked to the disabled Facebook profile. (D) 'User not found' shown for the linked Instagram account. (E) Re-appeal is denied once user agrees with platform's decision of violation. (F) Warning that the group might be disabled if more *admin violations* occur.

the privilege to resolve the restrictions by finding personal contacts who work at large social media platforms like Facebook. These findings show that instead of helping users, current appeal processes pushed them to further exploitation and oppression.

4.2.3 ***Anti-program and The Emerging Algospeak.*** Given the futility of the appeal process, participants were left on their own to devise ways to express their opinions while protecting themselves from future restrictions. They mostly relied on their intuition of how content moderation works. Many participants stopped posting publicly as they assumed public posts are more likely to be moderated. Page admins started adding English translations to their posts, assuming that their posts get flagged due to poor translation from Bengali to English by Facebook's built-in translation systems. Some group admins temporarily removed the admin who approved the flagged content, expecting that it would appease Facebook as they were warned that the group would be deleted in case of future admin violations (see Figure 2(F)). Some participants suspected they were restricted for using certain words, such as *Nazi*, *Hitler*, or *Taliban*. Hence, they either broke down such words or used codemixed Bengali and English letters and special symbols to write such words. Shafiq explained:

> *"After the Taliban regime came to power, Facebook started restricting anyone who mentioned Taliban. So people started writing Taliban in funny ways such as Tali*** or Tallu or 👏 ban to avoid restrictions. [The Bengali equivalent for clap is pronounced as Tali.]"*

Feenberg [49] defined such strategies as *anti-program* that users undertake to protest against the dominant forms of technology. Prior studies also report similar measures used by pro-Eating disorder communities to circumvent algorithmic restrictions [30, 53]. These findings demonstrate how the opaque moderation systems pushed users to subvert the injustices inflicted upon them with little to no resources.

4.3 Rethinking Moderation From Users' Perspective

We now present participants' recommendation to envision moderation by ingraining local context and values.

4.3.1 ***Making Community Standards Accessible.*** All participants demanded a concise and simplified outline of the community standards in local language and layman's terms. They wanted Facebook to provide a link to the community standards and specifics of the violation while accusing users of violating the community guidelines. Considering that most users might not have the time or skills to parse the community guidelines written in *legalese* and hidden behind multiple settings, our participants suggested creating short informative videos in local languages summarizing community standards. They recommended placing these videos either on the news feed or inserting them as short ads within regular videos, reels, and stories. They expected Facebook to communicate clearly what types of posts may lead to what restrictions and show example posts and associated restrictions. They also demanded access to the guidelines that moderators use for content moderation and more transparency about how Facebook handles mistakes in content moderation. Similar concerns about misclassification in content moderation have been also observed among the online gamers [79]. Ehsan highlighted the need for more transparency:

> *"Facebook should provide statistics on how many posts they moderate across different countries and languages. This will help us understand if there is any discrepancy in their moderation policies. They should also release how many posts were removed*

upon government request and how many for being reported by everyday users."

4.3.2 ***Moderate with Humans in the Loop.*** Some participants emphasized the need to train the underlying moderation algorithms on new datasets, believing historical data that feeds into these algorithms are more attuned to the Western norms of acceptable online speech. They demanded local representation in deciding content moderation policies, external review by local experts and everyday social media users, and tailored policies and tools for diverse communities in the Global South. Rashed elaborated:

"I heard about Facebook's research within several communities in India. But they should not consider India as a model of all South Asian countries because the contexts in India, Pakistan, Bangladesh, or Myanmar are very different."

Some participants suggested pro-active approaches to contain problematic content, including analyzing users' intent from their posts, comments, and prior conversations to inform them if their content might be potentially harmful and how they could avoid inadvertent use of hurtful language. They felt doing so might help the underlying models learn how users interact with different groups and prevent moral policing when they use coarse language within their close circle. Participants who were skeptical of AI's ability to understand different cultural contexts felt strongly about reviewing AI's decision by human moderators. They wanted the platforms to collaborate with local fact checking organizations who are well aware of local contexts. They also recommended that the platforms like Facebook should adopt new approaches to engage everyday users in content moderation which goes beyond passive reporting of harmful content. For example, they suggested that if a user's content is flagged by AI, Facebook should collect feedback from other users who interacted with the content because they might have a better understanding of the user's intent and the post's context. Moreover, to ensure that the human moderators have their powers in check, they advised to collect feedback from the users about the moderators' decision and take away moderation power from the moderators who repeatedly misjudge the content. Some participants also recommended involving group admins to resolve community standard violations within groups assuming they would know group members and the context better. However, others feared this might enforce admins' biases and would burden them with added invisible labor.

4.3.3 ***Moderate Within Context.*** Many participants were annoyed that Facebook restricted them without considering their longstanding record of good online behavior. They added that if an account is old enough and has no history of past offense, then Facebook should give a warning in case of first violation instead of enforcing restrictions immediately. They suggested that Facebook should consider the severity of the misconduct instead of rolling out the same restriction for just any violation. They expected the platform to decipher close friendships based on user interactions and not moderate casual interactions among close friends.

To reduce manipulation of the reporting feature, our participants advised that Facebook could collect more details from the reporting user about how and which parts of the reported content violate community standards. They expected Facebook to carefully check how old or recent the reported content is and if there is any anomaly in sudden outbursts of reporting. They also expected to receive a reference to the post that violated community standards and the details of violation. They proposed that Facebook should give users more time to either delete the flagged content or submit an appeal to overturn the decision of restriction. If users proactively delete the flagged content, they expected platforms to clear off the charge of violations against those users. Instead of placing punitive restrictions on users who violated community guidelines, they suggested that Facebook should invest more to repair, heal, and educate users. They advised to create educational courses on community standard violations and make it mandatory for users, who have violated the community guidelines. In line with the prior work by Schoenebeck et al. [125], participants suggested considering restorative justice approaches as a substitute to platform mediated punitive measures. Faruk described:

"Liberal societies have made a shift from punitive measures and adopted reformative justice approaches. If I call somebody names on Facebook, they can ask me to educate N users from my friend list about community standard violations or raise a certain amount of money online for a social cause."

Our participants suggested adopting alternative measures, such as coloring username in red or adding a badge of violation to the user profile. They also recommended adding a scoring system for users to promote good online behavior and erasing violation history if users maintain a streak of good online behavior. Participants also disapproved of banning Messenger, deleting users' Facebook accounts, pages, or groups without proper prior communication. Instead, they suggested to disable offending users' access to pages and accounts and give them an opportunity to download their data.

Participants also demanded several improvements to the appeal process. They wanted more time to process their emotions instead of being pushed to face the formalities of the appeal process immediately after a violation. They requested access to the flagged content to make cogent arguments and suggested that instead of deleting, Facebook could either blur the content or add a label indicating that the content is currently under review. They also wanted assistance to appeal the restrictions and expressed that Facebook should appoint customer service representatives to help users. These suggestions point that the current moderation processes fall short in meeting the users' needs and require substantive revisions.

5 DISCUSSION

Our findings show how Facebook's content moderation processes cause distress to Bangladeshi users for their non-Western ways of being. We first elaborate the colonial elements embedded in the design and implementation of current content moderation systems and discuss how the moderation infrastructure perpetuates *digital colonialism* by enforcing Western values upon users from the Global South countries like Bangladesh. We then discuss steps towards decolonial content moderation grounded within care by leveraging Bellacasa's work on care in ethics/politics, work/labor, and affect/affections [41].

5.1 Unraveling the Coloniality of Content Moderation

5.1.1 ***Community Standards.*** Like the colonial powers who socioeconomically exploited the colonized, our participants felt that Facebook and other Western social media platforms make huge profits by monetizing the data of users in the Global South, but do little to acknowledge and respect them and their diverse ways of being. All our participants, both living in Bangladesh and in the Global North, questioned the power imbalance in *who* gets to decide what is *appropriate* online conduct for users in Bangladesh and felt that they had no voice in shaping the community guidelines. They complained how inaccessible the community standards are for low-literate users and drew parallels with the historic subjugation of the low-literates in the British colonies. They compared the design and implementation of community standards to the top-down, power-driven, and command-based legal system that many colonized countries in the Global South, including Bangladesh, inherited from the colonial rulers [17]. They expressed that the current moderation approaches not only disregard the needs of Bangladeshi users, but also like colonial powers in the past, are downright oppressive, harassing, and exploitative. This is inline with the observation from Karanicolas [74], who has discussed how the platforms test new moderation policies on users in the Global South before launching them in the West.

Many participants also struggled with inconsistent moderation of religious posts and this made them doubt Facebook's integrity and neutrality to be an arbitrator. In fact, Facebook's negligence in moderating religiously charged hate speech and fake news has resulted in multiple instances of communal violence and attacks on religious minority and secular bloggers in Bangladesh [65, 142]. On one hand, Facebook's culturally and religiously ignorant moderation policies continue to silence users' voice against religious violence [44], and on the other hand, the underlying algorithms amplify extremist posts in full swing [77]. This pattern reflects the colonial legacy of using dominant culture and technology as a weapon for political and religious exploitation [109, 130].

5.1.2 ***Content Moderation.*** Despite social media platforms' use of "sophisticated" automated moderation techniques, the underlying algorithms still struggle to interpret harmful content in non-Western languages [27]. Not only research on automated moderation is highly concentrated in the Western regions, but also prioritizes the Western languages. For example, Facebook added its hate speech classifier for Bengali in 2020 and for violence and incitement in 2021 [38], much later than that of the Western languages, even though Bengali is the sixth largest spoken language in the world [45]. The harms resulting from the algorithmic inequity are further exacerbated when tools designed to tackle hate speech in the West are used in non-Western settings without much adaptation. For example, Facebook's mistranslation of the Arabic text *'good morning'* as *'attack them'* led to wrongful arrest [68] and its automated moderation tool could only detect 0.2% of the harmful content written in Afghan dialect [128]. Our participants clearly voiced concerns about such *algorithmic oppression* and latent biases in the content moderation pipeline. They blamed Facebook for using *faulty* datasets to train their models and believed that algorithms gave a slack when people used profanity in English, but took harsh measures when they swore in Bengali. These findings complement Davidson et al. [39], who showed systematic and substantial biases exist in the datasets that are widely used to train hate speech classifiers. Despite the shortcomings and opacity of the underlying AI algorithms, the platforms often justify AI as a solution to content moderation because its scalability allows them to grow further, situate themselves as an invisible infrastructure, and hide the politics of whose speech is allowed online [56, 57]. These systems, as aptly put by Cathy O'Neil [109], are *"weapons of math destruction"*, causing disproportionate harms to historically marginalized users in the Global South.

Even though some participants preferred human moderators to algorithms, systemic discrepancies in moderation infrastructure reduced their trust on them. Despite having most of their users based in the Global South, the Western social media platforms like Facebook hardly allocates enough resources to appoint the moderators who know local norms and languages. Prior work shows that the platforms often rely on English speaking moderators in the West to flag local content [136]. Even when the platforms recruit local moderators, they are paid extremely low wages [114] and are routinely subjected to psychological harm from the exposure to violent, graphic content [56].

There is also a huge power asymmetry between the platforms in the West and the third party moderator suppliers in the Global South. As the platforms exclusively control the software and technical infrastructure for content moderation, they can easily switch between different suppliers and find cheap labor elsewhere without having to bother about the poor working conditions of the moderators [3]. This mirrors the colonial exploitation of the non-Western labor force to enable the proper functioning of the Empire [108]. Moreover, due to the power asymmetries, human moderators are often pressured to label content conforming to populist views, else they risk sacrificing their wages [102]. As moderators feel obliged to finish their daily quota, they do not get enough time and context to assess the posts properly, making them susceptible to their own biases during moderation [135]. This results in moderators being blamed when things go wrong; in fact some of our participants also accused the moderators for pushing their own political and communal agenda. These findings point to *algorithmic exploitation* enabled by digital colonialism where Facebook and other social media platforms based in the West appropriate the cheap labor of the moderators based in the Global South to train their underlying algorithms while doing little to empower the moderators doing the ghost work [59].

5.1.3 ***Restrictions and Appeal Process.*** Our participants criticized Facebook's *paternalistic* attitude to penalize users without proper explanation. Following the colonial standards of legal justice system, current moderation systems deployed by most social media platforms prioritize punitive measures instead of reformative and restorative measures, and penalize the users without even letting them defend themselves. This inequity stems from the centuries-old colonial system that turned a blind eye to the right of the *accused* and their ability to afford the cost to defend themselves [51]. Colonial rulers posited the judge as a *"disinterested referee"* rather than *"an essential arm of power"* and thus, dismissed offenders' right to defense counsel as unnecessary [51]. However, within the cloak

of judicial neutrality ran deep-rooted biases against the people of color, indigenous people, and the colonized [42, 69]. Similarly, even though Facebook and other social media platforms claim to act as *neutral* [90], their power to decide which content to moderate as well as structural biases—both algorithmic and human—in moderation reproduce similar instances of colonial injustice for users in Bangladesh.

Moreover, the inadequate appeal process creates inequitable condition for users to seek remedy and support. Analogous to the judicial process, appeal in sociotechnical systems faces issues of asymmetric wealth, power, and access [147]. As our participants noted, Facebook has a flawed appeal process that is not equally offered to all users (inequitable), does not allow users to explain their point (oppressive), streteches beyond the length of the punitive restrictions (untimely), and neither clarifies the outcome of the appeal nor properly follows up with the appellant (arbitrary). Some of our participants bypassed the official process with the *"informal"* help of their acquaintances working at Facebook. This closely resembles the bureaucratic limbo introduced by the colonial rulers that is characterized by the poor delivery of services, nepotism, and corruption [63, 107].

In sum, our findings show how the elements of coloniality are deeply embedded in the current content moderation pipeline and: (1) impose oppressive and unjust restrictions on Bangladeshi users, and (2) center economic and political power in the hands of Facebook, a Western social media platform, that entitle it with unchecked power to regulate the press, speech, and online activities in foreign territories [81].

5.2 Towards a Decolonial Content Moderation

Using Bellacasa's framework of *"care"* [41], we propose decolonial content moderation as an ethical-political commitment to give voice and agency to the neglected users from the Global South communities like Bangladesh while going beyond the colonial epistemologies of universal ethics that produce such neglect.

5.2.1 ***Ethics/Politics.*** As private overlords of modern information infrastructure, social media platforms based in the West wield uninhibited political power to censor certain viewpoints and forms of speech with their proprietary black box algorithms [81]. Decolonial theories criticize the Western political hegemony to force a *universal ethics* in the name of global community standards and call for adopting pluriversal, intercultural ethics embedded in local values [46, 47, 94]. Decolonial processes necessitate *care* and *commitment* in designing technologies for, and with, the underserved communities [37]. Like Latour [83], Bellacasa rejects the design of technologies as *depoliticized* matters of fact [41, p. 18] and argues that the ethico-political meanings of care is not about normative moral obligations, but asking *"how to care"*, where care has the potential to subvert the status quo [41, p. 6].

Following Bellcasa's recommendation to go beyond normative moral obligations, the Western social media platforms like Facebook need to shift their paternalistic attitude and make an effort to understand the specific needs of each oppressed group. The platforms should engage in expansive intercultural dialogue with local stakeholders, policy makers, human rights organization, journalists, and media experts to integrate local norms and sensitivities in their content moderation policies. Instead of appointing a few *tokenistic* local experts, they should change and challenge the underlying systems of power [23]. They should strive for a process that produces *legitimate* actions and outcomes which are broadly accepted even by people who do not agree with all of them [13]. Moreover, they need to be more transparent about the internal working processes and power asymmetries shaping the community standards, and communicate their rationale behind imposing certain norms and how these norms would help the community [70]. They also need to make the community standards understandable to low-literate and non-English speaking users who are among the fastest growing user groups on social media.

5.2.2 ***Work/Labor.*** To subvert the exploitation by colonial division of work/labor that goes into content moderation—i.e., *esteemed* white collar jobs of highly paid engineers in the West and *low-status* blue collar jobs of low-wage moderators in the Global South—we need to develop a *critical* technical practice that recognizes power imbalances and implicit values in the design of AI systems [2, 97]. For this, the platforms need to do more than adopting *ethical AI* or *algorithmic fairness* approaches, especially since the mainstream notion of fairness and universal ethics based on the Western standards can reenact coloniality [1], lead to unethical outcomes for marginalized groups [104, 123], and benefit those already controlling the power and computing resources [116]. Bellacasa points that the rapid, innovation-driven imaginaries of popular AI systems present an objectified, market-dominated form of care [41, p. 23], which is in contrast with the care that posits itself as *"everything that is done"* to maintain, continue, and repair the world to live as well as possible [41, p. 161]. Thus, the platforms need to critically reflect on the work—i.e., the choices, assumptions, dataset selection, and fairness considerations—that shape predictive content moderation systems [96].

First, there should be fundamental changes in design practices to connect the AI practitioners with the organizational and institutional realities, and more importantly, with the constraints and needs of users on the ground. Second, the platforms need to adopt fairness-aware computing approaches to learn how historical biases shape moderation outcomes [89] and enact participatory design practices by allowing stakeholders, who are directly impacted by moderation outcomes, to define *fairness* in their own terms [73]. The platforms should use community-engaged approaches rather than putting the burden of representation on a single individual [148]. Moreover, the platforms need to be more transparent about their automated moderation systems, for example, what training and test data they use, false positive and false negative rates of the underlying models, and how the models encode *appropriateness and fairness* in different contexts and geographies.

Social media platforms rarely make visible the *care work* that the low-paid moderators in the Global South undertake to maintain the platforms. Bellacasa notes how technological design often devalues everyday human labor as *ordinary* and enforces the colonial divisions of labor [41, p. 54], much like we see in content moderation. She explains that *"care can turn into moral pressure for workers when they rightfully try to preserve their affective engagement from exploitations of waged labor"* [41, p. 5]. To put an end to the exploitation of the low-paid moderators, the platforms need to deal with

the issues around moderators' wage, career growth, and working conditions. For example, even when the platforms outsource moderation to third party contractors, they should keep the working environment in check. Besides, as our participants were divided between their preferences for human moderators, the platforms need to disclose more meta data about their moderators in different geographic regions, for example, how they are recruited, their language and cultural expertise, what training and guidelines they received, and what fraction of local content is moderated by them. Moreover, to reduce moderators' biases, the platforms should get each post reviewed by multiple reviewers [110] and properly communicate to users how many moderators wanted to keep/delete the post and how the disagreement was resolved [13]. The platforms should also develop new tools that enable the moderators to access the moderation rules with examples of violating content [72], access violators' past history, and share that information with other moderators [26]. Moreover, algorithmic flagging can be used to highlight the problematic segments in a content [140] or to automatically distill useful information from an article [100] and surface high quality comments in a post so that moderators can take multiple perspectives into account [112]. However, such tools should be designed and deployed with care so that they empower the moderators against the colonial appropriation of digital labor.

In adddition, the platforms need to acknowledge the *free labor* that users undertake by voluntarily reporting harmful content and tactfully approach when to take action against legitimate user reporting and when to dismiss trolls weaponizing the feature to harass other users. For example, the platforms can use data analytics to identify reliable flaggers [79] and acknowledge and empower them, in ways similar to YouTube's trusted flagger program [153]. The platforms need to inform general users about how reporting works, follow up with the reporting user about what action they took, and offer more clarity to the reported user about how they differentiate between reporting by legitimate users and trolls [78, 79]. As our participants suggested, the platforms should provide new features that would enable the reporting user to easily annotate problematic parts of the content and describe how it violates community standards.

5.2.3 ***Affect/Affections.*** Bellacasa frames affective dimensions of care as taking responsibility for other's well-being [41, p. 162]. The platforms need to embed care in each phase of the content moderation pipeline to support users, especially since our participants noted the dramatic impact of unjust and unfair restrictions on their mental health, well-being, social interactions, and work life. As a first step, the platforms need to *"listen with care"* to the concerns of the users in the Global South. Bellacasa discusses that listening with care is an active political process that shapes whose and what concerns are ratified [41, p. 58]. Second, a decolonial approach aims to reorient content moderation towards repairing, educating, and sustaining communities [129]. The platforms should shift from punitive restrictions to more reformative, reflective, and restorative measures, at least for users who violated community standards the first time. For example, the platforms can proactively analyze the content before users' posts are made available to others and *educate* them how the content might violate community guidelines when applicable [70, 151]. Prior research shows that proper explanations decrease the likelihood of future content violations [71]. Moreover, when the platforms take extra time and care to explain their rationale behind moderation, they also increase the perceived legitimacy of the moderation outcome [110].

Moreover, the platforms need to provide both mental health care (e.g., clinical care, resilience training) and technological support (e.g., tools to blur graphic content or see them in grayscale) to protect the moderators from the psychological harms of content moderation [137]. They need to consider the severity of harmful content to balance the work load across multiple moderators to reduce overburden [124]. The platforms should also offer mental health support to everyday users who report problematic content routinely and thereby might be exposed to disturbing materials.

In addition, contestability—the opportunity to meaningfully challenge the platforms' moderation decisions [64]—is important to design a fair and just moderation system. The platforms need to give users more time to process and appeal restrictions. They need to provide a fair process that allows users to express their arguments and provide feedback on the usefulness of the appeal process. Besides, to relieve users from the burden of understanding complex standards and the emotional fatigue that accompanies the appeal process, the platforms should provide emotional support to the accused users [147] and provide resources to formulate effective arguments, like structured rubrics and appeal etiquette [72, 147]. Such affective and non-paternalistic care-based measures are needed not only to decolonize content moderation, but also to preserve human dignity.

6 LIMITATIONS AND CONCLUSION

This paper unpacks how coloniality in Facebook's current content moderation systems amplify power relations, center the Western norms, and erase minoritized expressions of users from the Global South communities like Bangladesh. We use Bellacasa' three-dimensional framework of care to propose steps towards a fundamentally decolonial content moderation infrastructure that would center the voices of the socially and culturally diverse social media users in the Global South.

Our study has some shortcomings. Apart from the inherent limitations of small sample size in qualitative research, our sample study population is skewed towards educated, affluent, urban, and male Facebook users in Bangladesh. Our sample had only one female participant despite our sincere efforts to recruit them (e.g., advertising the study through the Facebook account of a female researcher and interviews led by a female researcher). While we do not know the reasons why female Facebook users did not show interest to participate in the study, we assume it might be because of social stigma around receiving penalties from Facebook, which many of our male participants also noted and this could be heightened for female users in a deeply patriarchal society. We also acknowledge that the experiences of Bangladeshi Facebook users and their needs and aspirations related to content moderation may not be generalized to diverse user groups in the Global South. Future work should identify the commonalities and differences among users in different geographies and of varying socioeconomic status, urbanity, and gender identities. Moreover, our work only focuses on the moderation practices deployed by Facebook. Even though there

are similarities in the exclusionary content moderation pipelines of various Western social media platforms and users' experiences of them [75, 151], more work is needed to investigate the effects of content moderation on historically marginalized users across different Western social media platforms. While our work takes the important first steps towards reimagining a decolonial content moderation infrastructure that centers the voices of users in the Global South, who are critically underrepresented in the current research advances–more work is needed to empirically evaluate the merits of the proposed measures across diverse Global South communities.

REFERENCES

[1] Rachel Adams. 2021. Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews* 46, 1-2 (2021), 176–197.

[2] Philip E. Agre. 2006. Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*. Psychology Press, UK, 131–157.

[3] Sana Ahmad and Martin Krzywdzinski. 2022. Moderating in Obscurity: How Indian Content Moderators Work in Global Content Moderation Value Chains. In *Digital Work in the Planetary Market*. The MIT Press, Cambridge, MA.

[4] Emmanuel Akinwotu. 2021. *Facebook's role in Myanmar and Ethiopia under new scrutiny*. The Guardian. Retrieved September 1, 2022 from https://www.theguardian.com/technology/2021/oct/07/facebooks-role-in-myanmar-and-ethiopia-under-new-scrutiny

[5] Syed Mustafa Ali. 2016. A Brief Introduction to Decolonial Computing. *XRDS* 22, 4 (jun 2016), 16–21.

[6] Ali Alkhatib. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 95, 9 pages.

[7] Adriana Alvarado Garcia, Juan F. Maestre, Manuhuia Barcham, Marilyn Iriarte, Marisol Wong-Villacres, Oscar A Lemus, Palak Dudani, Pedro Reynolds-Cuéllar, Ruotong Wang, and Teresa Cerratto Pargman. 2021. Decolonial Pathways: Our Manifesto for a Decolonizing Agenda in HCI Research and Design. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 10, 9 pages.

[8] New America. 2021. Case Study: Reddit. Retrieved November 15, 2022 from https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/case-study-reddit/

[9] Ángel Díaz and Laura Hecht-Felella. 2021. Double Standards in Social Media Content Moderation. Retrieved November 15, 2022 from https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

[10] Samaya Anjum. 2022. *Concerted attacks against Bangladeshi activists on Facebook*. Global Voices. Retrieved July 11, 2022 from https://globalvoices.org/2022/02/08/concerted-attacks-against-bangladeshi-activists-on-facebook/

[11] Ahmed Ansari. 2019. Decolonizing Design through the Perspectives of Cosmological Others: Arguing for an Ontological Turn in Design Research and Practice. *XRDS* 26, 2 (nov 2019), 16–19.

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

[13] Shubham Atreja, Libby Hemphill, and Paul Resnick. 2022. What is the Will of the People? Moderation Preferences for Misinformation. Retrieved August 26, 2022 from https://arxiv.org/abs/2202.00799

[14] Laima Augustaitis, Leland A. Merrill, Kristi E Gamarel, and Oliver L. Haimson. 2021. Online Transgender Health Information Seeking: Facilitators, Barriers, and Future Directions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 205, 14 pages.

[15] Kagonya Awori, Nicola J. Bidwell, Tigist Sherwaga Hussan, Satinder Gill, and Silvia Lindtner. 2016. Decolonising Technology Design. In *Proceedings of the First African Conference on Human Computer Interaction* (Nairobi, Kenya) *(AfriCHI'16)*. Association for Computing Machinery, New York, NY, USA, 226–228.

[16] Madeline Balaam, Rob Comber, Rachel E. Clarke, Charles Windlin, Anna Ståhl, Kristina Höök, and Geraldine Fitzpatrick. 2019. Emotion Work in Experience-Centered Design. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12.

[17] Hussain M. Fazlul Bari. 2019. Evolution of the criminal justice system in Bangladesh: colonial legacies, trends and issues. *Commonwealth Law Bulletin* 45, 1 (2019), 25–46.

[18] Seyla Benhabib. 1992. *Situating the self: Gender, community, and postmodernism in contemporary ethics*. Psychology Press, East Sussex, UK.

[19] Cynthia L. Bennett, Daniela K. Rosner, and Alex S. Taylor. 2020. The Care Work of Access. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15.

[20] Gurminder K Bhambra, Dalia Gebrial, and Kerem Nişancıoğlu. 2018. *Decolonising the university*. Pluto Press, Las Vegas, NV, USA.

[21] Sam Biddle, Paulo Victor Ribeiro, and Tatiana Dias. 2020. *TikTok told moderators to suppress posts by "ugly" people and the poor to attract new users*. The Intercept. Retrieved September 8, 2022 from https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/

[22] Nicola J Bidwell. 2016. Moving the centre to design social media in rural Africa. *AI & society* 31, 1 (2016), 51–77.

[23] Abeba Birhane and Olivia Guest. 2021. Towards Decolonising Computational Sciences. *Kvinder, Køn & Forskning* 29 (02 2021), 60–73.

[24] Oversight Board. 2021. Oversight Board demands more transparency from Facebook. Retrieved November 15, 2022 from https://www.oversightboard.com/news/215139350722703-oversight-board-demands-more-transparency-from-facebook/

[25] Elena Botella. 2019. *TikTok Admits It Suppressed Videos by Disabled, Queer, and Fat Creators*. Slate. Retrieved September 8, 2022 from https://slate.com/technology/2019/12/tiktok-disabled-users-videos-suppressed.html

[26] Jie Cai and Donghee Yvette Wohn. 2021. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 410 (oct 2021), 25 pages.

[27] Katie Canales. 2021. *Facebook's AI moderation reportedly can't interpret many languages, leaving users in some countries more susceptible to harmful posts*. Insider. Retrieved July 25, 2022 from https://www.businessinsider.com/facebook-content-moderation-ai-cant-speak-all-languages-2021-9

[28] Transparency Center. 2021. Takedown experience. Retrieved November 15, 2022 from https://transparency.fb.com/en-gb/enforcement/taking-action/taking-down-violating-content/

[29] Twitter Help Center. 2021. Our range of enforcement options. Retrieved November 15, 2022 from https://help.twitter.com/en/rules-and-policies/enforcement-options

[30] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1201–1213.

[31] Baijayanti Chatterjee. 2020. Ecology and Imperium: State Formation in Early Colonial Bengal c. 1765–1800. *Indian Historical Review* 47, 2 (2020), 263–281.

[32] Christine L. Cook, Aashka Patel, and Donghee Yvette Wohn. 2021. Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms. *Frontiers in Human Dynamics* 3 (2021), 8 pages.

[33] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806.

[34] Nick Couldry and Ulises A Mejias. 2019. Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media* 20, 4 (2019), 336–349.

[35] Nick Couldry and Ulises A Mejias. 2020. *The costs of connection: How data are colonizing human life and appropriating it for capitalism*. Oxford University Press, Oxford, UK.

[36] John W. Creswell and Dana L. Miller. 2000. Determining Validity in Qualitative Inquiry. *Theory Into Practice* 39, 3 (2000), 124–130.

[37] Cristiano Codeiro Cruz. 2021. Decolonizing philosophy of technology: Learning from bottom-up and top-down approaches to decolonial technical design. *Philosophy & Technology* 34, 4 (2021), 1847–1881.

[38] Elizabeth Culliford and Brad Heath. 2021. *Language Gaps in Facebook's Content Moderation System Allowed Abusive Posts on Platform: Report*. The Wire. Retrieved August 11, 2022 from https://thewire.in/tech/facebook-content-moderation-language-gap-abusive-posts

[39] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 25–35.

[40] Giovanni De Gregorio. 2020. Democratising online content moderation: A constitutional framework. *Computer Law & Security Review* 36 (2020), 105374.

[41] María Puig de la Bellacasa. 2017. *Matters of Care: Speculative Ethics in More Than Human Worlds.* University of Minnesota Press, Minnesota, MN, USA.

[42] Randle C DeFalco and Frédéric Mégret. 2019. The invisibility of race at the ICC: lessons from the US criminal justice system. *London Review of International Law* 7, 1 (06 2019), 55–87.

[43] Paul Dourish and Scott D. Mainwaring. 2012. Ubicomp's Colonial Impulse. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, Pennsylvania) *(UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 133–142.

[44] Vera Eidelman, Adeline Lee, and Fikayo Walter-Johnson. 2021. *Time and Again, Social Media Giants Get Content Moderation Wrong: Silencing Speech about Al-Aqsa Mosque is Just the Latest Example.* ACLU. Retrieved August 10, 2022 from https://www.aclu.org/news/free-speech/time-and-again-social-media-giants-get-content-moderation-wrong-silencing-speech-about-al-aqsa-mosque-is-just-the-latest-example

[45] Chad Emery. 2022. *The 33 Most Spoken Languages in the World (2021).* Langoly. Retrieved August 18, 2022 from https://www.langoly.com/most-spoken-languages/

[46] Arturo Escobar. 2011. Sustainability: Design for the Pluriverse. *Development* 54 (06 2011), 137–140.

[47] Charles Ess. 2006. Ethical pluralism and global information ethics. *Ethics and Information Technology* 8 (11 2006), 215–226.

[48] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14.

[49] Andrew Feenberg. 2017. Critical theory of technology and STS. *Thesis Eleven* 138, 1 (2017), 3–12.

[50] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 40 (may 2020), 28 pages.

[51] Laurie S. Fulton. 1989. The Right to Counsel Clause of the Sixth Amendment. *American Criminal Law Review* 26, 4 (1989), 1599–1616.

[52] Akriti Gaur. 2020. *Moderate Globally Impact Locally: Tackling Social Media's Hate Speech Problem in India.* Yale Law School. Retrieved August 10, 2022 from https://law.yale.edu/moderate-globally-impact-locally-tackling-social-medias-hate-speech-problem-india

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

[54] Ysabel Gerrard and Helen Thornham. 2020. Content moderation: Social media's sexist assemblages. *New Media & Society* 22, 7 (2020), 1266–1286.

[55] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 1–5.

[56] Tarleton Gillespie. 2021. *Custodians of the Internet.* Yale University Press, New Haven, CT, USA.

[57] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 1–15.

[58] Colin M Gray and Elizabeth Boling. 2016. Inscribing ethics and values in designs for learning: a problematic. *Educational technology research and development* 64, 5 (2016), 969–1001.

[59] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (illustrated edition ed.). Mariner Books, Boston, MA, USA.

[60] David Greene, Paige Collings, and Christoph Schmon. 2022. *Online Platforms Should Stop Partnering with Government Agencies to Remove Content.* Electronic Frontier Foundation. Retrieved September 1, 2022 from https://www.eff.org/deeplinks/2022/08/online-platforms-should-stop-partnering-government-agencies-remove-content

[61] Ramón Grosfoguel. 2011. Decolonizing post-colonial studies and paradigms of political-economy: Transmodernity, decolonial thinking, and global coloniality. *Transmodernity: journal of peripheral cultural production of the luso-hispanic world* 1, 1 (2011), 38 pages.

[62] Greg Guest, Kathleen M. MacQueen, and Emily E. Namey. 2012. *Applied thematic analysis.* SAGE, Thousand Oaks, CA.

[63] Akhil Gupta. 1995. Blurred Boundaries: The Discourse of Corruption, the Culture of Politics, and the Imagined State. *American Ethnologist* 22, 2 (1995), 375–402.

[64] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 466 (oct 2021), 35 pages.

[65] Mubashar Hasan, Geoffrey Macdonald, and Hui Hui Ooi. 2022. *How Facebook Fuels Religious Violence.* Foreign Policy. Retrieved August 10, 2022 from https://foreignpolicy.com/2022/02/04/facebook-tech-moderation-violence-bangladesh-religion/

[66] Rebecca Heilweil. 2021. YouTube's newest content moderation stat, briefly explained. Retrieved November 15, 2022 from https://www.vox.com/recode/2021/4/6/22368809/youtube-violative-view-rate-content-moderation-guidelines-spam-hate-speech

[67] YouTube Help. 2021. Harmful or dangerous content policies. Retrieved November 15, 2022 from https://support.google.com/youtube/answer/2801964?hl=en

[68] Alex Hern. 2017. *Facebook translates 'good morning' into 'attack them', leading to arrest.* The Guardian. Retrieved August 17, 2022 from https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

[69] Tanveer Rashid Jeewa and Jatheen Bhima. 2021. Discriminatory Language: A Remnant of Colonial Oppression. *Constitutional Court Review* 11, 1 (2021), 323–339.

[70] 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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 192 (nov 2019), 33 pages.

[71] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 150 (nov 2019), 27 pages.

[72] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proc. ACM Hum.-Comput. Interact.* 4, GROUP, Article 17 (jan 2020), 35 pages.

[73] Christopher Jung, Michael J. Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. 2019. Eliciting and Enforcing Subjective Individual Fairness.

[74] Michael Karanicolas. 2020. *The Countries Where Democracy Is Most Fragile Are Test Subjects for Platforms' Content Moderation Policies.* Slate. Retrieved August 10, 2022 from https://slate.com/technology/2020/11/global-south-facebook-misinformation-content-moderation-policies.html

[75] Michael Karanicolas. 2020. *Moderate globally, impact locally: A series on content moderation in the Global South.* Yale Law School. Retrieved August 25, 2022 from https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/moderate-globally-impact-locally-series-content-moderation-global-south

[76] Naveena Karusala, Aditya Vishwanath, Arkadeep Kumar, Aman Mangal, and Neha Kumar. 2017. Care as a resource in underserved learning environments. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[77] Karen Kornbluh. 2022. *Disinformation, Radicalization, and Algorithmic Amplification: What Steps Can Congress Take?* Just Security. Retrieved August 10, 2022 from https://www.justsecurity.org/79995/disinformation-radicalization-and-algorithmic-amplification-what-steps-can-congress-take/

[78] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 334 (oct 2021), 21 pages.

[79] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 437, 12 pages.

[80] Chris Köver and Markus Reuter. 2019. *TikTok curbed reach for people with disabilities.* Netzpolitik.org. Retrieved September 8, 2022 from https://netzpolitik.org/2019/discrimination-tiktok-curbed-reach-for-people-with-disabilities/

[81] Michael Kwet. 2019. Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class* 60, 4 (Apr 2019), 1–20.

[82] Ganaele Langlois, Greg Elmer, Fenwick McKelvey, and Zachary Devereaux. 2009. Networked Publics: The Double Articulation of Code and Politics on Facebook. *Canadian Journal of Communication* 34 (08 2009).

[83] Bruno Latour. 1987. *Science in action: How to follow scientists and engineers through society.* Harvard university press, Cambridge, MA, USA.

[84] Shaimaa Lazem, Danilo Giglitto, Makuochi Samuel Nkwo, Hafeni Mthoko, Jessica Upani, and Anicia Peters. 2021. Challenges and paradoxes in decolonising HCI: A critical discussion. *Computer Supported Cooperative Work (CSCW)* 31, 0 (2021), 1–38.

[85] Jessa Lingel. 2021. *The Gentrification of the Internet.* University of California Press, Oakland, CA.

[86] María Lugones. 2007. Heterosexualism and the Colonial / Modern Gender System. *Hypatia* 22, 1 (2007), 186–209.

[87] Kim Lyons. 2020. *Facebook reportedly bracing for US election chaos with tools designed for 'at-risk' countries.* The Verge. Retrieved September 11, 2022 from https://www.theverge.com/2020/10/25/21533352/facebook-us-election-chaos-at-risk-countries-trump

[88] Renkai Ma and Yubo Kou. 2022. "I'm not sure what difference is between their content and mine, other than the person itself": A Study of Fairness Perception of Content Moderation on YouTube. *PACM on Human Computer Interaction* 6, CSCW, Article 425 (11 2022), 28 pages.

[89] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased

Data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 349–358.

[90] Andrew Marantz. 2020. *Why Facebook Can't Fix Itself?* The New Yorker. Retrieved August 17, 2022 from https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself

[91] Amanda Meng, Carl DiSalvo, and Ellen Zegura. 2019. Collaborative Data Work Towards a Caring Democracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 42 (nov 2019), 23 pages.

[92] Meta. 2021. Counting strikes. Retrieved November 15, 2022 from https://transparency.fb.com/en-gb/enforcement/taking-action/counting-strikes/

[93] Walter Mignolo. 2011. *The darker side of western modernity: Global futures, decolonial options.* Duke University Press, Durham, NC, USA.

[94] Walter D. Mignolo. 2012. *Local Histories/Global Designs: Coloniality, Subaltern Knowledges, and Border Thinking.* Princeton University Press, Princeton, NJ, USA.

[95] Walter D Mignolo and Catherine E Walsh. 2018. *On decoloniality: Concepts, analytics, praxis.* Duke University Press, Durham, NC, United.

[96] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (mar 2021), 141–163.

[97] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33, 4 (2020), 659–684.

[98] Maria D. Molina and S. Shyam Sundar. 2022. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society* 0, 0 (2022), 1–19.

[99] Michelle Murphy. 2015. Unsettling care: Troubling transnational itineraries of care in feminist health practices. *Social studies of science* 45, 5 (2015), 717–737.

[100] Kevin K. Nam and Mark S. Ackerman. 2007. Arkose: Reusing Informal Information from Online Discussions. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '07)*. Association for Computing Machinery, New York, NY, USA, 137–146.

[101] Ashis Nandy. 1989. *Intimate Enemy:Loss and Recovery of Self Under Colonialism.* Oxford University Press, Oxford, UK.

[102] Aliide Naylor. 2021. *Underpaid Workers Are Being Forced to Train Biased AI on Mechanical Turk.* VICE. Retrieved August 13, 2022 from https://www.vice.com/en/article/88apnv/underpaid-workers-are-being-forced-to-train-biased-ai-on-mechanical-turk

[103] Sabelo J. Ndlovu-Gatsheni. 2015. Decoloniality as the Future of Africa. *History Compass* 13, 10 (2015), 485–496.

[104] Helen Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (2001), 120–119.

[105] Nel Noddings. 2012. The language of care ethics. *Knowledge Quest* 40, 5 (2012), 52.

[106] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[107] Dele Olowu. 1988. Bureaucratic Morality in Africa. *International Political Science Review / Revue internationale de science politique* 9, 3 (1988), 215–229.

[108] Gail Omvedt. 1973. Towards a Theory of Colonialism. *Insurgent Sociologist* 3, 3 (1973), 1–24.

[109] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway books, Portland, OR, USA.

[110] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 82 (apr 2022), 31 pages.

[111] Naresh R Pandit. 1996. The creation of theory: A recent application of the grounded theory method. *The qualitative report* 2, 4 (1996), 1–15.

[112] 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* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1114–1125.

[113] Sachin R Pendse, Daniel Nkemelu, Nicola J Bidwell, Sushrut Jadhav, Soumitra Pathare, Munmun De Choudhury, and Neha Kumar. 2022. From Treatment to Healing:Envisioning a Decolonial Digital Mental Health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 548, 23 pages.

[114] Billy Perrigo. 2022. *Inside Facebook's African Sweatshop.* TIME. Retrieved August 13, 2022 from https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

[115] Twitter Developer Platform. 2022. Safety Tools: Bodyguard. Retrieved September 1, 2022 from https://developer.twitter.com/en/community/toolbox/bodyguard

[116] Aníbal Quijano. 2000. Coloniality of Power and Eurocentrism in Latin America. *International Sociology* 15, 2 (2000), 215–232.

[117] Aníbal Quijano. 2007. Coloniality and modernity/rationality. *Cultural studies* 21, 2-3 (2007), 168–178.

[118] Rema Rajeshwari. 2019. *Mob Lynching and Social Media.* Yale Journal of International Affairs. Retrieved September 1, 2022 from https://www.yalejournal.org/publications/mob-lynching-and-social-media

[119] Reddit. 2021. Reddit Content Policy. Retrieved November 15, 2022 from https://www.redditinc.com/policies/content-policy

[120] Martin J. Riedl, Kelsey N. Whipple, and Ryan Wallace. 2021. Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others. *Information, Communication & Society* 0, 0 (2021), 1–18.

[121] Sarah T. Roberts. 2018. Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday* 23, 3 (Mar 2018).

[122] Jacqueline Rowe. 2022. Marginalized Language and The Content Moderation Challenge. Retrieved September 1, 2022 from https://www.gp-digital.org/marginalised-languages-and-the-content-moderation-challenge/

[123] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-Imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 315–328.

[124] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 368 (oct 2021), 33 pages.

[125] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from justice theories to support targets of online harassment. *New Media & Society* 23, 5 (2021), 1278–1300.

[126] Sarita Schoenebeck, Carol F. Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair After Online Harassment. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 2 (apr 2021), 18 pages.

[127] Tristan Schultz, Danah Abdulla, Ahmed Ansari, Ece Canlı, Mahmoud Keshavarz, Matthew Kiem, Luiza Prado de O Martins, and Pedro JS Vieira de Oliveira. 2018. What is at stake with decolonizing design? A roundtable. *Design and Culture* 10, 1 (2018), 81–101.

[128] Mark Scott. 2021. *Facebook did little to moderate posts in the world's most violent countries.* Politico. Retrieved August 17, 2022 from https://www.politico.eu/article/facebook-content-moderation-posts-wars-afghanistan-middle-east-arabic/

[129] Eugenia Siapera. 2022. AI Content Moderation, Racism and (de) Coloniality. *International Journal of Bullying Prevention* 4, 1 (2022), 55–65.

[130] Anjuans Simmons. 2015. *Technology Colonialism.* Model View Culture. Retrieved August 10, 2022 from https://modelviewculture.com/pieces/technology-colonialism

[131] Nathaniel Sirlin, Ziv Epstein, Antonio A. Arechar, and David G. Rand. 2021. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School Misinformation Review* 2, 6 (Dec. 2021), 13 pages. https://doi.org/10.37016/mr-2020-83

[132] Dhevy Sivaprakasam and Raman Jit Singh Chima. 2022. OPINION: Content takedowns on social media facilitate censorship in Asia. Retrieved September 1, 2022 from https://news.trust.org/item/20220426081524-d7orq

[133] CalvinJohn Smiley and David Fakunle. 2016. From "brute" to "thug:" The demonization and criminalization of unarmed Black male victims in America. *Journal of human behavior in the social environment* 26, 3-4 (2016), 350–366.

[134] Linda Tuhiwai Smith. 2021. *Decolonizing methodologies: Research and indigenous peoples.* Bloomsbury Publishing, London, UK.

[135] Olivia Solon. 2017. *Underpaid and overburdened: the life of a Facebook moderator.* The Guardian. Retrieved August 15, 2022 from https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content

[136] Steve Stecklow. 2018. *Why Facebook is losing the war on hate speech in Myanmar.* Reuters. Retrieved July 25, 2022 from https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

[137] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 341, 14 pages.

[138] Judith Sutz. 2021. Thinking otherwise: The ambiguous role of technological imaginaries in development processes. Retrieved September 9, 2022 from https://www.youtube.com/watch?v=BaH4OCKQgGo

[139] Nicolas P. Suzor. 2019. *Lawless: The Secret Rules That Govern our Digital Lives.* Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108666428

[140] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. 2021. Effects of Algorithmic Flagging on Fairness: Quasi-Experimental Evidence from Wikipedia.

Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 56 (apr 2021), 27 pages.

[141] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2022. (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society* 0, 0 (2022), 1–22.

[142] Ishaan Tharoor. 2016. *These Bangladeshi bloggers were murdered by Islamist extremists. Here are some of their writings.* The Washington Post. Retrieved August 16, 2022 from https://www.washingtonpost.com/news/worldviews/wp/2016/04/29/these-bangladeshi-bloggers-were-murdered-by-islamist-extremists-here-are-some-of-their-writings/

[143] Austin Toombs, Shad Gross, Shaowen Bardzell, and Jeffrey Bardzell. 2017. From empathy to care: a feminist care ethics perspective on long-term researcher–participant relations. *Interacting with Computers* 29, 1 (2017), 45–57.

[144] Austin L. Toombs, Shaowen Bardzell, and Jeffrey Bardzell. 2015. The Proper Care and Feeding of Hackerspaces: Care Ethics and Cultures of Making. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 629–638.

[145] Emily Tseng, Mehrnaz Sabet, Rosanna Bellini, Harkiran Kaur Sodhi, Thomas Ristenpart, and Nicola Dell. 2022. Care Infrastructures for Digital Security in Intimate Partner Violence. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 123, 20 pages.

[146] Eve Tuck and K Wayne Yang. 2021. Decolonization is not a metaphor. *Tabula Rasa* 1, 1 (2021), 61–111.

[147] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 167 (oct 2020), 22 pages.

[148] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 318 (oct 2021), 28 pages.

[149] Sofia Villenas. 1996. The colonizer/colonized Chicana ethnographer: Identity, marginalization, and co-optation in the field. *Harvard educational review* 66, 4 (1996), 711–732.

[150] Human Rights Watch. 2018. Bangladesh: Crackdown on Social Media. Retrieved September 1, 2022 from https://www.hrw.org/news/2018/10/19/bangladesh-crackdown-social-media

[151] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.

[152] Marisol Wong-Villacres, Adriana Alvarado Garcia, and Javier Tibau. 2020. Reflections from the Classroom and Beyond: Imagining a Decolonized HCI Education. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–14.

[153] YouTube. 2020. About the YouTube Trusted Flagger program. Retrieved August 27, 2022 from https://support.google.com/youtube/answer/7554338

[154] Bingjie Yu, Joseph Seering, Katta Spiel, and Leon Watts. 2020. "Taking Care of a Fruit Tree": Nurturing as a Layer of Concern in Online Community Moderation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9.