Sustained Harm Over Time and Space Limits the External Function of Online Counterpublics for American Muslims

NILOUFAR SALEHI, University of California, Berkeley, USA
ROYA PAKZAD, Taraaz, USA
NAZITA LAJEVARDI, Michigan State University, USA
MARIAM ASAD, Sassafras Tech Collective, USA

Social media platforms are celebrated for their capacity to empower those with marginalized or disenfranchised identities and support them to create counterpublics. We focus on one such group, Muslim Americans, and ask how visible Muslim Americans, such as journalists, activists, and aspiring politicians, use social media to craft counter-narratives, reclaim control of their stories, and mitigate the harm directed at them. Through a series of 19 interviews, we found that visible Muslim Americans' ability to craft and sustain counter narratives is largely hampered by sustained online harm (e.g. harassment). We found that these public figures were harmed repeatedly over long periods of time and through the weaponization of platform affordances such as replying, tagging, and hashtag takeovers, as well as the weaponization of their gender and identity. Our findings shed light on the serious limitations of social media to provide a safe platform for counterpublics to engage externally with wider publics. Finally, we discuss the limitations of content moderation as the dominant framework for addressing harm online and suggest alternative paths forward based on restorative and transformative justice.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Online harm, harassment, content moderation, Muslim Americans

ACM Reference Format:
Niloufar Salehi, Roya Pakzad, Nazita Lajevardi, and Mariam Asad. 2023. Sustained Harm Over Time and Space Limits the External Function of Online Counterpublics for American Muslims. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 93 (April 2023), 24 pages. https://doi.org/10.1145/3579526

1 INTRODUCTION

Social media platforms have been applauded for being at the forefront of progress towards an open and inclusive public sphere - a place where people can gather, share ideas, and debate [23, 148]. However, the experiences of marginalized people in these spaces remain complicated [137]. Marginalized people's participation in online spaces opens them to frequent and severe harms such as physical threats, stalking, harassment, and sexual abuse [160]. At the same time, HCI researchers have studied how marginalized people benefit from online safe spaces that afford a degree of escape from harm and can support discourse, identity formation, and mutual aid [137]. These safe spaces are reminiscent of Nancy Fraser's formulation of *subaltern counterpublics*, or private spaces for discourse that marginalized groups have created outside of traditional public

Authors' addresses: Niloufar Salehi, University of California, Berkeley, Berkeley, California, USA, nsalehi@berkeley.edu; Roya Pakzad, Taraaz, Santa Cruz, California, USA, rpakzad@taraazresearch.org; Nazita Lajevardi, Michigan State University, Michigan, USA, nazita@msu.edu; Mariam Asad, Sassafras Tech Collective, Ann Arbor, Michigan, USA, mariam@sassafras.coop.

2573-0142/2023/4-ART93
https://doi.org/10.1145/3579526

spaces. Here, we focus on the role that online platforms play in the *external function* of subaltern counterpublics *i.e.* persuading larger society that the ideas formulated by the counterpublic are valid [148].

To explore this, we study how Muslim Americans who are highly visible – *e.g.* journalists, activists, and public officials – engage with broader publics online, are harmed in the process, and use different strategies to mitigate those harms. Muslim Americans face unique and severe forms of online abuse, such as harassment and hate speech on social media sites [33, 85]. After a July 2020 audit, Facebook found that the social media giant had itself "created an atmosphere where Muslims feel under siege" on the platform [49]. In response to these high risk conditions, Muslims in the U.S. have largely reduced their visibility online [84], suffered heightened anxiety [131], disengaged politically [122], and have experienced negative health effects [14]. At the same time, however, these conditions have also encouraged Muslim Americans to mobilize: more have run for office [151], some have become more vocal online [84], and there has been more organizing among a new base of Muslim leaders, particularly within progressive and radical movements [47]. These efforts can be explained through Fraser and Squire's definitions of counterpublics [61, 148].

In this paper, we study what happens when Muslim American counterpublics engage online with broader, dominant publics. We conducted semi-structured interviews with 19 self-identified Muslims living in the United States who have public personas and an active presence on social media platforms. We found that participants rely heavily on their social media networks to engage with other publics, respond to and challenge dominant narratives that demonize Muslims, and to broadcast and uplift counter-narratives. In doing so, participants frequently faced abuse and harassment, which were often sustained over long periods of time and space, weaponized platform affordances, and more severely affected people with overlapping marginalized identities (*e.g.* racialized Muslims, women and non-binary/gender non-conforming Muslims).

Despite these risks, participants continued to engage with online audiences and platforms, often while trying to leverage platform affordances to enact more agency over their safety and wellbeing, such as by curating online audiences and developing safety plans. We found that participants rarely, if ever, reported safety violations to platforms due to a lack of faith in their efficacy, the laborious reporting process, and the burden of reliving their traumas. In some cases, participants reported that online safety mechanisms – *e.g.* blocking, reporting – resulted in *increased* abuse. We end with a discussion of potential directions for addressing online harm and building safer and more just online spaces based on principles of restorative and transformative justice [21, 116, 139, 162].

2 RELATED WORK

Our paper draws on two main areas of scholarship. First, we discuss research on publics, counterpublics, and how they form and operate online. Second, we discuss research on online harm, such as harassment, particularly focusing on their impacts on marginalized people. Our work builds on recent HCI scholarship that studies and designs for the needs of Muslims and other faith-based communities [5, 6, 117, 129, 130, 152].

2.1 Online Publics and Counterpublics

Theoretical work about public spheres have heavily shaped the study of media and democracy, and in particular work on publics and counterpublics [11, 148]. These theories originate from the conception of the singular *bourgeois public sphere* by Habermas [75], together with Fraser's later response [61]. For Habermas, the public sphere is a theatre in modern societies where all citizens can engage in political participation through the medium of talk. This concept of a public sphere imagines a physical or mediated space, distinct from the state, where citizens can share information and discuss and deliberate on shared issues. For Habermas, an active public sphere is necessary for

a functioning democracy. Unfortunately, most of the time, participation in public discourse is not accessible to all people. In fact, the rise of the bourgeois public sphere that Habermas documents in the eighteenth and nineteenth centuries in Europe, coincides with the African slave trade and dehumanization of people of color [13]. Therefore, Habermas failed to acknowledge how women and people of color were largely left out of the ideal public sphere of the Enlightenment era [148].

In 1990, Nancy Frasier linked work on the Habermasian public sphere with the preoccupations of the subaltern school of postcolonial historiography [61]. Frasier coined the term *subaltern counterpublics*, as "parallel discursive arenas where members of subordinated social groups invent and circulate counterdiscourses, which in turn permit them to formulate oppositional interpretations of their identities, interests, and needs" [56, 61]. Fraser argues that subaltern counterpublics need safer, separate spaces to discuss their interests outside of the interference of the dominant public [61]. Even if access to the public sphere is theoretically guaranteed to all, social norms and state surveillance can instill fear in marginalized people that their participation in counterpublics might be met with indifference, or worse, violence [148]. While Habermas posits that members of the public should bracket or leave behind their identities when they enter the public sphere, Fraser argues that such bracketing does not remove power relations, but merely obscures them [61]. Like Fraser, many contemporary scholars have articulated multiple, coexisting counterpublics composed mainly of marginalized people [25, 132, 148]. Coexisting counterpublics may be differentiated by group characteristics or identities, such as ethnicity, race, gender, or religion. Take for instance an immigrant support group formed in reaction to exclusionary politics of the dominant public and/or the state [148]. The conceptual move away from the Habermasian ideal of a single public sphere, allows for the recognition of the political struggles and innovations of marginalized people outside of public spaces dominated by white cisgender bourgeois men [148].

One area of ambiguity in this literature is around what makes a counterpublic, "counter"? Is any gathering and discussion among marginalized people a counterpublic, or is the creation of new counter-narratives that critique the dominant order needed? Does a marginalized group need to take actions with the goal of transforming the dominant order to be recognized as a counterpublic? To answer this question, Squires contributes a framework based on Black public spheres in the U.S. [148]. Squires recognizes three different types of responses that a marginalized public sphere might produce given their particular political, social, and cultural conditions: *enclave*, *counterpublic*, and *satellite*. A public might *enclave* itself –hiding ideas that are counter to the dominant view in order to protect itself– while engaging in internal deliberation and development of those ideas and strategies. The second type of public is a *counterpublic* that can engage in debate with wider publics in order to test ideas and potentially take collective action. Finally, a *satellite* public seeks separation for reasons other than oppression and may engage in wider discourse from time to time. Squires posits that the form a public takes is not only a reaction to outside oppression, but is also shaped by the public's internal relations, culture, and access to resources, all conditions that may change over time [148].

With more discourse moving online particularly through social media and online forums, subaltern counterpublics gained an important new technological medium for connecting and finding common cause. As many media scholars have documented, online spaces often allow for the emergence of new alternatives to hegemonic social relations and power structures [48, 98]. Relatedly, HCI researchers have documented protected, safe spaces created by marginalized people where they can engage in discussion, identity formation, and mutual support (e.g. [7, 16, 40, 51, 77, 81, 86, 94, 108, 116, 130, 136, 137, 149, 155, 163]). This body of work contributes mechanisms that can create safe spaces online, such as anonymity [7], privacy settings, the creation and enforcement of community standards [128], content warnings, and limited, controlled membership [77]. While

these safe online spaces have many benefits for marginalized people, they can also themselves be infiltrated and targeted [77, 137].

While having access to safe, protected spaces for discourse is necessary, in this paper, we focus on the *external function of counterpublics*: persuading the larger society that the counterpublic's claims are valid [54, 148]. For instance, Felski describes the institutional mechanisms and sites which created the avenues for the dissemination of feminist thought, such as health clinics, political action groups, and bookstores [54]. Similarly, Dawson highlights the importance of various Black institutions in the success of Black public spheres such as the Black press, popular music, and the church [41]. Therefore, the political success of a counterpublic is impacted by the institutions it is able to form, its ties to dominant political actors, and its ability to create effective vehicles for publicity [148]. Historically, publicity by a counterpublic has been facilitated through independent media sources and distribution channels [61]. Social media can provide one such avenue.

Jackson et al. rely on examples such as #BlackLivesMatter, #FastTailedGirls, and #MeToo [63], to argue that various forms of hashtag activism can serve as offense or defense mechanisms to attract attention to a social and political issue or change harmful narratives against marginalized groups [87, 98]. As disembodied public spaces that are not racially marked, platforms such as Twitter are perceived as white. Kuo explains that in such an environment, racial justice activist hashtags such as #Solidarityisforwhitewomen and #NotYourAsianSideKick demonstrate injustice, re-frame discourse, and/or promote policy change while at the same time establishing grounds for participation and building collective identity [99]. Research has also shown how the use of hashtags such as #CanYouHearUsNow and #MuslimsReportStuff help American Muslims exert control on their narratives online [124]. Such activism sometimes involves the use of humor and satire, which has been shown to be a powerful tool for creating more interaction in countering hate speech [15]. In addition, some studies have examined how hashtag hijacking has been used by marginalized communities to change the dominant or original narrative of a hashtag (e.g. #IstopIslam, #myNYPD) [88].

Despite the promise of online platforms to facilitate the external function of counterpublics, in this work we found that visible Muslim Americans' capacity to make use of this avenue was severly limited by pervasive online harms that stretched through time and space.

2.2 Online Harms and Platform Moderation

Prior work in HCI has studied how people, particularly marginalized people, are harmed online, including through hatespeech [29, 80], violent imagery [107], misinformation [55], harassment [19, 20, 121, 125], and bullying [12, 100, 147]. Online harms cause emotional distress and jeopardize the physical safety of their targets [29, 158]. Researchers have also studied the gendered forms of online harm [110, 158] and how gendered online harm has offline impact [127]. One study found that online harassment is prevalent for young women and that platform affordances increase opportunities for harassment [159]. Similarly, people of color face myriad harms online [46, 73, 97, 120], and social media reliance has been shown to increase support for hostile anti-Muslim policies, in particular [104].

Content moderation is the dominant mechanism that platforms use to address a myriad of online harms [68, 133, 142]. Some platforms also utilize volunteer moderators who mostly focus on interpersonal harm, and can set rules, remove content, and ban people [142]. However, existing forms of content moderation have been criticized for failing to remove a range of severe harms such as posts promoting alt-right views and revenge porn [31, 119, 157]. Recently, machine learning tools have begun to play an important role in automating content moderation by flagging and remove violating content [27]. However, a lack of nuanced or cultural understandings can result

in failures such as for detecting hate speech [91]. Research has shown how negative sentiment around Islam and Muslims are actually built into these automated systems [1].

Content moderation practices have had limited success in helping marginalized people. Many content moderation practices are based on user reports which favor majority norms [35]. But even when content that harms marginalized people is reported, platforms fail to act on it. In one study researchers at the Center for Countering Digital Hate reported 530 posts which contained disturbing, bigoted, and dehumanizing content targeting Muslim people. The social media companies (Facebook, Instagram, TikTok, Twitter, and Youtube) failed to act on 89% of the posts [59]. Together these posts were viewed at least 25 million times. The same research center has shown that social media platforms fail to act on antisemitism [58] and misogynist abuse [60].

On the other hand, people with marginalized identities are more likely to have their content removed [164]. One study found that transgender people had their content removed when it was related to their identity, contained adult content (despite following site guides), or critical of dominant groups and Black people had their content removed when they spoke about racism or racial justice [79]. Overall, content related to expressing their marginalized identity was removed despite following site policies or being within moderation gray areas [79]. Other research has found that content about racism is frequently flagged and removed as hate speech [74] even when the context is a Black person sharing a racist incident that had happened to them [89]. Black content creators have said that TikTok hides their posts especially if they post about race or Black Lives Matter [64] and Trans users have accused TikTok of censoring trans content [36].

In this paper, we study what happens when Muslim counterpublics in the U.S. engage externally and how online harms limit their ability to do so. In the next section, we will provide background on the current state of Muslims in the U.S.

3 BACKGROUND: MUSLIMS IN THE U.S.

Muslims became central to the U.S. sociopolitical discourse in the aftermath of the September 11, 2001 terrorist attacks. In the years since, Muslims have experienced immense hostility in the societal and political realms [90, 102, 105, 106, 144]. But, Islamophobia is nothing new; it dates back to the antebellum South when the first Muslims arrived as forcibly enslaved people [17]. Historically, Muslims have been constructed at odds with the "west," and have been stereotyped as backwards, undemocratic, and uncivilized [134]. Notwithstanding the immense diversity of customs, histories, traditions, and languages among the global and domestic Muslim populations, popular tropes of Islam are traditionally linked to violence, misogyny, barbarism, and intolerance [17, 52, 101, 103, 123, 134]. Thus, despite being a diverse religious group, Muslims in the U.S. have been racially coded, identified, named, and categorized [38, 143], with their cultural and religious values and practices routinely criticized and posed in opposition to democratic norms and principles [37, 39, 143]. Moreover, anti-Muslim attitudes are pervasive and growing: in 2017, the number of anti-Muslim hate groups rose for the third straight year and the Council on American Islamic Relations (CAIR) recorded 2,599 anti-Muslim bias incidents taking place across the nation [101]. Especially worrisome for the how Muslims fare in everyday sociopolitical life in the U.S. is the large degree to which hostility has also manifested on online platforms; for instance, the 2018 election in the U.S. saw coordinated activities on Twitter to track Muslim candidates and utilize Islamophobia to mobilize voters against them [126].

Researchers have begun to explore the effects of increasing stigmatization on Muslims themselves. Muslims report high levels of discrimination [105, 122], and those with more experiences of discrimination are more likely to hold a higher sense of "stigma consciousness" [141], or awareness of their low sociopolitical positioning. Assaults against Muslims in 2017 surpassed post-9/11 figures

in 2001, [93] and in 2017, 75% of Muslims agreed that there is a lot of discrimination against their group [43].

Veiled Muslim women, in particular, are uniquely positioned to experience harm. Research has found that Muslim women face discrimination in the labor market [2, 42, 65], are rated as less intelligent and attractive and are more linked to tropes of terror [76, 109], and face immense barriers to healthcare [146, 154]. Those who wear the hijab are especially vulnerable, because the headscarf serves as a visible marker of their religious identity. In the wake of the 2016 presidential election, the Southern Poverty Law Center noted that numerous veiled Muslim women were physically attacked and grabbed by their headscarves [26]. Given that Muslim women are especially visible and that they have experienced physical and unprovoked offline harm by merely existing, they may be experiencing similar, if not more violent, forms of abuse online, where attackers have the privilege of being anonymous and rarely held accountable.

Facing this rising tide of hostility, some Muslims have identified their visibility as putting them at greater risk. For example, one month after the 2016 election, in response to rising anti-Muslim hate crimes and Muslim women indicating that they felt unsafe in public, Imam Abdullah Antepli of Raleigh, North Carolina, recommended to women that under the extraordinary circumstances under which Muslim Americans now live, they may require extraordinary measures to protect their safety, including taking off the hijab, at least temporarily. Imam Omar Suleiman of Irving, Texas likewise echoed this advice when considering hijabi women's fears and advised them to make practical changes to protect their safety, such as wearing a hoodie instead of a hijab if they felt in danger [24, 38, 145].

This climate, moreover, has negatively affected U.S. Muslims and Arabs who at least temporarily reduced their visibility in public spaces, both online on Twitter and offline in public spaces. In the wake of major discriminatory campaign events throughout the 2016 presidential campaign, Arab and Muslim Twitter users reduced their visibility online [84]. And, some scholarship has found that in the wake of the Trump's 2017 Executive Order restricting travel from majority Muslim countries, also commonly referred to as the "Muslim Ban," U.S. Muslims reported that the ensuing protests in their defense highlighted their visibility and made them feel more likely to endure physical harm [101].

In reaction to dominant narratives that demonize and dehumanize them, some Muslim Americans have become more visible and vocal and have engaged in counterpublic activities. In this paper, we study what happens when these counterpublics attempt to use online social media platforms to engage with dominant publics.

4 STUDY DESIGN

Our research was guided by an over-arching question: how do visible Muslim Americans use social media platforms to engage with and share their own narratives within wider dominant publics? Early in the course of this research we learned about the severe impact of online harm, such as harassment and hate speech on our participants, and included studying the impact and nature of that harm as a second research question.

To answer these research questions, we conducted semi-structured interviews with 19 self-identified Muslims living in the United States who have public personas and an active presence on social media platforms. Below, we outline our recruitment and interview procedure, discuss ethical considerations in carrying out this research, and describe the limitations we encountered.

4.1 Data Collection

4.1.1 Interview Recruitment: We used purposeful sampling to recruit interview participants. By querying public databases of Muslim policy and advocacy groups, activists, Muslim journalists,

academics, and political candidates, we aggregated a list of 85 American Muslims and organizations active online. The phrase "active online" was defined as a user who possessed a public account on at least one social media platform and frequently used that account for their personal and professional communications such as posting content, replying to other content, following other users, or reporting. Due to our research interest in the relationship between Muslim American counterpublics and the dominant American public, our initial list consists of people with a public platform who are active on social media in English. We divided our list into the following categories: activist or organizer; politician; Muslim policy and advocacy group's representative; academic; journalist; and other public figures including comedians and bloggers. We emailed all 85 individuals on the list and invited them to a 45 to 60 minute zoom interview. In our invitation email we outlined the purpose of our research, our funding sources, and the researchers' names and affiliations. In addition we attached the interview questions and an interview recording and data collection consent form to ensure that a participant have enough time to read the interview questions and learn about the purpose of their engagement. Our study and interview questions were approved by the Institutional Board Review at our university.

4.1.2 Interview Procedure: Nineteen individuals accepted our invitation for interviews. From June to August 2020, we interviewed them in 45 to 60 minute video interviews through Zoom. These interviews were recorded, and were semi-structured. Our interviews followed four procedures. Upon arrival, the participant was introduced to the researcher(s) and the study. Any questions about the consent form and audio-recording were addressed by the researcher(s). The interviews began with a short introduction of the participant about their profession, the social media platforms they use most often, and the ways they track their online engagement. The second part of the interview sought to understand participants' self-identification as a Muslim on a social media platform, the harms they faced, and the strategies they have taken to protect themselves and to exert agency over their representation online. We followed this with questions about potential limitations of social media platforms' design features and policies with respect to protecting the participants and their communities. We concluded the interviews by inquiring about the participants' recommendations for social media platforms and asked if there have been any other experiences that they would like to share with us. We believe that our participants' lived experience as Muslims in America is likely to be shaped by their demographic backgrounds [138]. As such, following the interviews, we sent a form to inquire about participants demographic information including gender, race and ethnicity, and the title they would like to be associated with in the paper. We offered the participants a USD $50 gift-card as compensation for their time. Table 1 indicates our participants' demographic backgrounds.

4.2 Data Analysis

We hired two research assistants to manually transcribe the interviews. We then analyzed the data in an iterative process of interpretative qualitative coding [111]. Before beginning the coding procedure, we developed a shared understanding of the interviews' main themes and developed a preliminary codebook. Two of the co-authors coded all 19 interviews line by line adhering to an open coding process by taking inductive approaches. They used the research group's Slack channel to address uncertainties and refine codes. In addition, this subset of researchers reviewed each other's coding schemes to ensure consistency. Occasionally, they also wrote analytical memos based on their observations during the interviews and coding process. In the end, the codebook consisted of three levels of codes including 20 top-codes, 74 subcodes, and 17 sub-subcodes. These researchers later used a virtual whiteboard to share and discuss the study's main themes and findings.

Participant ID	Gender	Racial-Ethnic Group	Profession
P1	Male	Arab American	Journalist
P2	Male	South Asian American	Public Speaker
P3	Female	Central Asian American	Former Political Candidate
P4	Male	Latinx American	Researcher
P5	Male	Middle Eastern American	Academic
P6	Female	White American	Former Political Candidate
P7	Female	Arab American	Former Political Candidate
P8	Female	Arab American	Executive Director of a Muslims Policy Organization
P9	Female	Middle Eastern American	Academic
P10	Female	Arab American	Academic
P11	Female	Middle Eastern American	Blogger and Activist
P12	Female	Arab American	Blogger
P13	Male	Arab American	Former Political Candidate
P4	Male	South Asian American	Communications Director of a Muslims Advocacy Group
P15	Female	Arab American	Journalist
P16	Female	South Asian American	Journalist
P17	Male	South Asian	Politician
P18	Female	Arab American	Executive Director of a Muslims Advocacy Organization
P19	Female	Arab American	Journalist

Table 1. Demographic information of the research participants

4.3 Researchers' Reflexive Statement and Ethical Considerations

Given the sensitivity of our research topic, we acknowledge that discussing the lived experiences of racism, sexism, Islamophobia, and other forms of online harms can be challenging for participants [9]. We addressed this concern in different ways: we used semi-structured interviews as a method to ensure that participants have more control over the conversation. In addition, we extended invitations to participants to review our work at the end of each interview, as well as in post-interview surveys. It is important to note that the research approach was an example of member-research because all members of the research group (including research assistants) self-identify as Muslim [3]. As such, we have some degree of understanding about the harassment and discrimination that Muslim Americans face. This also helped us to establish trust with our participants such that they felt comfortable sharing their experiences with us.

4.4 Limitations

According to a 2017 PEW Survey, 20% of Muslim Americans are racially Black [115]. Despite our efforts, we were not successful in recruiting any Black Muslims for this study. In addition, none of our participants identified as LGBTQ+, despite increasing numbers of Muslims indicating their LGBTQ+ status in surveys [150]. Black and queer Muslims are heavily marginalized in the U.S. and the absence of their voices limits our ability to paint a comprehensive picture of the nature of harm that visible Muslims face. In particular, the intersections of marginalized identities for Muslim Americans increases their risks and vulnerabilities to both external harm (by the dominant public) and internal harm (by other members of the counterpublic). Additionally, we would like to reflect on the fact that our sampling strategy specifically reached out to Muslim Americans who had a visible, public profile online. We speculate that having this kind of public presence might be even more costly and dangerous for more heavily marginalized Muslims in America which made it difficult for us to recruit such people. In other words, as some of our participants acknowledged, despite their marginalized religious and ethnic identities, our participants still benefited from other

types of privilege such as being white-presenting or having access to resources through class privilege or institutional networks. Future work can study how overlapping identities and levels of access can impact a counterpublic's external function, as well as the role that internal power relations play in shaping who 'speaks' for a counterpublic [148].

5 RESULTS

Our participants developed strategies to take control of their own narratives. They worked to build an audience and used their platforms to educate and craft counternarratives to dominant narratives that demonized them. Counternarratives are stories that people tell and live by in resistance to dominant cultural narratives [8]. Our participants saw value in working to create and promote counternarratives about their identities and lives as Muslim Americans. However, they were also acutely aware of the limitations of this approach given lack of organizational resources and limitations on the volume of their speech.

Our participants thought carefully about their audiences, how to engage with them, and how to build community. Some participants limited their audiences and focused on community building, others took a more public approach and engaged in public online debates. These participants described activities that were close to the external functions described for counterpublics [148]. In other words, even if they didn't necessarily believe that they could change the minds of the people they were talking to directly, they saw public debates as an opportunity to perform for the "silent watchers" and transform culture. One participant described how they didn't see their public facing work as necessarily stopping white supremacists from enacting violence, but more as an effort to change the culture that condoned that violence and made it possible:

> "If one is thinking that putting out mass messages is going to counter hate, and maybe stop a Brendan Terrance, a Dylan Roof, or a Robert Bowers, I think we need to be a little more circumspect and realistic. If we are saying, though, it's going to shift the national temperature which would make the overall climate of hatred toned down and in turn, turn down the social permissibility of violent action, then that's a different thing altogether." (P4)

When people of marginalized identities, such as Muslim Americans, become visible online, they are harmed [4]. Our participants described the chilling effect that this had on members of their communities:

> "I definitely see that people are more scared. People are more concerned for their safety. Even when people do talk to me I have some people who say like can you not put my name if you want to put me in a story or don't put my picture because they're afraid of a backlash. They're afraid that people online [...] are going to look up their names look up their locations and threaten them. And that has happened unfortunately many times. So it definitely makes it harder [...]. You have people who choose not to come online at all. They don't want to make Facebook accounts they don't want to make Twitter accounts. They don't want to engage with their local media. They don't want to talk to media. And so we're now putting pressure on these people and we're forcing them to." (P15)

Our participants described online harm that is pervasive, organized, and meant to intimidate and silence them. They talked about how social media harassment led to letters written to employers, Chancellors, and university Presidents to silence them. How whole websites (e.g. canary mission) were set up to catalog and monitor their activities and make them hyper-visible and vulnerable. Finally, they talked about the chilling effect that this harm ultimately has on their ability to engage in public discourse and shape their own narratives.

> "Being public as a Muslim, at least as an American Muslim in today's political climate [...] is putting people in a vulnerable position, where even if you want to be vocally Muslim - to say you are Muslim - you are potentially putting yourself at harm, or putting yourself at risk for reading Islamophobic comments, or getting death threats, or having people say things about them, or threaten them, whatever it might be. I think that's the difficult side of Muslim Americans being on social media and you know, being public." (P12)

In this section we describe the ways that visible Muslim Americans were harmed online. Understanding the nature and characteristics of the harm that visible marginalized people face, helps us better understand the role that social media plays in both providing a platform for and limiting the voices of marginalized people. Understanding the harm is also the first step to effectively address it.

5.1 Sustained harm over long periods of time and (virtual) space.

One of the main characteristics of the harm that our participants described was sustained harm over long periods of time and (virtual) space. Here, by space we mean virtual spaces in social media platforms such as posts, private messages, email, comments sections, and tags, as well as across multiple platforms. Prior research has found simlar characteristics in the harms experienced by female journalists [72]. Sustained harm over space and time – whether that be harassment, hate speech, or threats of physical violence – has two main effects: it wears the recipient down and it makes their social platforms unusable.

> "It's like, months and months of just, like [being targeted and attacked] and it sends you into really suicidal depression. It's not a joke at all." (P8)

When harm is stretched over time and space, it becomes significantly more damaging and intimidating. The temporal aspects of the harm were particularly salient to our participants. They described harm that happened over long periods of time, sometimes months, as well as harm that happened quickly and repeatedly. These patterns meant that it was difficult for the recipient to find space to distance themselves from the harm and to heal. Additionally, our participants described harm that happens consistently over long periods of time as intimidating and debilitating. The implicit message is that someone took the time to find information about them, write messages, or edit images and create memes in order to harm them. For instance, one participant described knowing the fact that a person was paying constant attention to them as particularly threatening:

> "That one individual for example who was tweeting about me continuously [...] it was so continuous. And it was the length. It just seemed a bit more threatening because the person was real." (P16)

Similarly, when harm was stretched across multiple virtual spaces participants felt hyper-visible and accessible. Sometimes this escalated to doxxing or the threat of physical harm:

> "For me, when something's on social media, it's very distant, it's still distant and anonymous. But once somebody calls you on your work phone, emails you, takes the time to find your email at work– which, it's easy to find, but still they're going to that extent– leaving you a message, writing you emails, it's not very much more for them to find out where you live." (P10)

Sustained harm over time and space makes social media platforms unusable for the targets. Participants described how the sheer amount of notifications, tags, and mentions overrode the basic functionality of their social media accounts. This led many to be forced to close their accounts for some amount of time. Consistent with prior literature we found that this also leads to people

loosing their channels for social support. One participant described how they stopped checking their notifications because of the magnitude of the harm:

> "Since [our successful] campaign went trending, I received at least a Tweet every thirty seconds for 2 and a half weeks straight. There were graphic images, there were death threats [...] all of these things for two and a half weeks. So for me, I couldn't see anyone who had been tweeting because I wouldn't check my notifications. The second I would scroll to the top it would hit 20+, 20+." (P11)

These findings highlight the importance of considering time and space when designing ways to address online harm. The temporal and spatial aspects of online harm create a challenge for dominant content moderation frameworks. Content moderation generally considers individual pieces of content at the moment they are reported, which becomes less effective when harm happens quickly and repeatedly and when it accumulates over long periods of time. Additionally, the main available enforcement mechanism for users in a content moderation framework is reporting content that is against the rules. However, because the harm is spread over time and space using content moderation reporting functions becomes very labor intensive and almost impossible. As one participant described:

> "One of these accounts – if they target you, and they have sometimes millions of followers – you're kind of dead, you're just dead. You're going to be harassed for months. And I don't know what to do about it. People just get off Twitter. This keeps happening to people I know, and it's very hard to figure out what to do. I mean, they're blocking and reporting, but you almost need like a full-time staff to help you fight it, and you're usually on your own." (P8)

Overall, for visible Muslim Americans what made online harm particularly damaging was sustained harm over time and space. The length and consistency of such harm made using the basic functionalities of their social media accounts impossible, from posting content, to checking their notifications, to reporting the harm itself. This often led to people becoming silenced or forced to close their accounts.

5.2 Weaponizing platform affordances: reply, tag, quote tweet, hashtag, report.

The second major attribute of the harm that our participants described was the weaponizing of platform affordances. Our participants described how the people who harmed them took advantage of almost any platform affordance to do so. This included requests, tagging, hashtag takeovers, replying, quote tweeting, blocking, reporting, and even influencing algorithms used by the platform.

Affordances are the actionable properties between the world and an actor [66]. Social media sites afford posting, tagging, responding, etc. Prior research has introduced the concept of disaffordances, or perceptual cues with actions that are blocked or constrained as well as dysaffordances, or objects that require the person to misidentify themselves to be able to act on [34, 161]. Designers have been called on to take account of disaffordances and dysaffordances in order to prevent systematic exclusion of marginalized people [34]. While we agree with the importance of these calls, our participants were blocked or constrained from platform affordances not because of disaffordances in the design of the platform, but because of the intentional weaponization of those or other affordances.

Weaponized affordances are affordances that are used with the express purpose of harming someone. For instance, prior research has shown the weaponization of location tracking applications for intimate partner violence [62]. When platform affordances are weaponized they are often used as tools to overwhelm the target by harming them over space and time. This in turn often makes it

impossible for the target to make use of the platform. For instance, multiple participants talked about how tagging was used as a tool to harm them or people they knew.

> "A friend getting totally buried and overwhelmed with notifications, where they're just being tagged in hateful text online, like on Twitter or something like that. Or they're being tagged in very violent or pornographic images that are meant to bully or intimidate or harass." (P9)

Another example of weaponizing affordances is hashtag takeovers. One of our participants explained how #CAIR was initially created by CAIR, the Council on American-Islamic Relations, a nonprofit, grassroots civil rights and advocacy organization, as a mechanism for dialog and raising awareness about Muslim issues in the U.S. However, the hashtag was quickly taken over by people who used it to spread Islamophobic hate-speech, including violent and disturbing messages. This participant described how people struggled to regain their voice in shaping the narrative:

> "You know it's funny I would say that some of our people do that by attempting to reclaim the hashtag but there's just so much junk in it that you know [...] you're either creating a new hashtag or reclaiming an existing hashtag but you're constantly dealing with the possibility that there's just so much more keyboard power on the other side." (P18)

In all of the cases that we found, power enabled the weaponizing of affordances. In some cases that meant the labor intensive "keyboard power" of repeatedly tagging, messaging, or reporting someone or taking back a hashtag. We argue that in addition to considering platform disaffordances and dysaffordances, designers should also consider and study how affordances might be weaponized and design safeguards to stop that harm.

5.3 Weaponizing gender and identity.

The final characteristic of the harm that our participants described was the weaponizing of their gender and their Muslim identity. Our participants described their online presence as being under increased scrutiny because of their marginalized identity, their visibly Muslim profile, or their name that was perceived as foreign. The harm our participants described was also heavily dependent on their gender. One participant spoke to the specific ways this gendered harm manifested for visibly Muslim people:

> "Women are sexualized. Women are seen as oppressed [...] men are seen as scary and dangerous and so the ways in which the harassment can manifest can be different. Different in content not in severity." (P18)

Overall, we found that people in our study had very different conceptions and attitudes toward the harm that they experienced based on their gender. This is in line with prior work that has shown that women are more likely to view online harassment as a major problem [45]. They are also more likely to be targets of it. In one study, researchers set up fake online user accounts. They found that female-sounding names were 25 times more likely to receive threatening or sexually explicit messages [112]. Accounts that seemed to belong to women received 100 such messages (threatening or sexually explicit messages) on average every day, compared to 3.7 for accounts that seemed to belong to men [112]. Similarly, the women we talked to were more frequently targeted with sexual harassment and gender based harm.

> "There's also a lot of sexual harassment as well, I would guess some of that is tied to being Muslim because there is still a sexualization and orientalization that goes into it, and also the perversiveness of like, sending like, a dick pic to a hijabi" (P11)

One woman participant described the severity and gendered nature of the harm she experienced, particularly as she gained visibility:

> "And that's when I would get attacked quite aggressively, after I was publishing articles, and most of them, they were attacks that were direct emails to my work account. Those were the most vulgar. They were very gendered. They were very xenophobic, so the usual 'go back to your country,' 'you're a terrorist,' and then there were these threats of rape." (P10)

Identity based harm is not limited to Islamophobic hate speech. Our participants experienced harm at the intersection of their gender and identity. For instance, some Muslim women reported being targets of harassment by other Muslims for not conforming to social or cultural expectations. One participant explained how she struggled to make sense of this harm because it fell outside of dominant Islamophobic attacks:

> "I didn't think about it initially as something that could be Islamophobic or whatever, but it is targeted for me because I'm a [...] visibly Muslim woman, is usually from other Muslims. And it's usually like, you know, 'haram police' or it's somebody who is being critical of something I've shared, that they don't think is within the guidelines of our faith. Or it's something about the way I dressed, or it's something about the way I act [...] that's way more common that I've experienced but I guess in my mind I don't put it in the same folder, I guess, as Islamophobic" (P12)

This participant described the complexities of being simultaneously harmed by the dominant culture and by members of the same marginalized group, or "insider harm" [137]. This experience is not unique to Muslims, researchers have also documented the complexities of experiencing "cultural betrayal" by Black women who have been harmed by Black men [69]. Such harms are associated with the break down of trust and the development of cultural betrayal trauma [69]. Finally, acknowledging and describing these complex forms of harm is complicated for us, the authors, as it necessitates the discussion of within-group conflict and harm within a dominant academic culture with ingrained Islamophobia [116].

The weaponizing of gender and identity shows the importance of analyzing harm within the complex cultural and political context that it is happening. Meaningfully addressing harm without doing so is not possible. In this section we described the three main characteristics of the harm that our participants experienced: 1) sustained harm over long periods of time and space, 2) weaponizing platform affordances, and 3) weaponizing gender and identity. In the next section we will describe the strategies that our participants used when faced with this harm and the constraints they faced.

5.4 Gendered strategies for protection

We found that the strategies that people used for their safety and protection were also gendered. The men we talked to mostly took an approach of ignoring the harm and the offenders whereas women described detailed plans for protecting themselves. The reason for this difference is both in the types of harm that men and women receive as well as their perceived vulnerabilities. The Muslim men we interviewed were mostly worried about protecting their reputations, whereas the Muslim women we interviewed were worried about protecting their physical and psychological safety.

> "I would say that you know with women sort of it's about safety for men it's the opposite it's being reported to law enforcement. They think women are oppressed so it's not unimaginable that once people see you as oppressed they further pile on to that, and men are seen as dangerous" (P18)

The men we talked to in this study mostly described ignoring and not engaging in the face of online harm.

> "You can attack me all you want. I don't care that much." (P13)

This may be due to what each person perceives as the magnitude of risks to their safety and well-being, and may also be related to cultural expectations that men perform strength in public. Even when one participant was threatened physically and had to take protective actions, he downplayed it as being expected:

> "When I ran, I wasn't tracking my own social media as much as the team was. And we just got enough death threats that folks were like, 'You probably should take a bodyguard.' and you know, it's not like I didn't know it was coming. I hate to say it but it is part in parcel for running for office and being Muslim. But also running a prominent campaign in a [anonymized] office." (P13)

The women we talked to described being afraid of potential harm, particularly physical harm.

> "And so my big concern was that somebody was gonna come over to my house and do something, and I have children. And I'm worried that they would come to the school. And you always know where I am when I'm teaching." (P10)

Women were also more likely to have established plans for their safety and the safety of their family and friends that they shared with us. For instance, one participant said that they never tweet about an event until they have left the location and another had communicated with friends and family about what information they can and cannot post about them online. Most of the women we talked to had a detailed plan for blocking, documenting, limiting their audience, and asking friends and the community for help, as well as hiring a bodyguard if needed.

> "I don't post photos of my family online on Facebook and Twitter and so I'm very hyper-aware and I think I've done a lot of preventative things that has avoided that hate and the trolling translating from the online world to the physical world [...] I think a lot of that has to do with me being proactive and me having the privilege of having organizational resources and support. So I think that has been extremely helpful and making sure that the online stays online." (P15)

These gendered differences in both perceptions of risk and strategies for protection shed light on who is most likely to be removed from public discourse and how. This knowledge can be used as guidance for how to design mechanisms to address and prevent the harm.

5.5 Content moderation was considered mostly useless, sometimes harmful

To our surprise, almost all participants told us that they do not find platform content moderation useful for the types of harm that they face. The participants described abandoning content moderation reporting tools after many failed attempts. One participant told us that content moderation was sometimes weaponized against them. They described multiple instances where their content was taken down after mass reporting attacks and that they then had to go through a long process to reinstate their content.

Our participants described experiences of reporting content, waiting for a long time, and ultimately hearing back that the content did not violate any rules.

> "I was blocking, muting, and reporting in all of these cases. Specific to Facebook, I've reported so many people who have directly slandered me, said things that weren't true, said things that put me in danger, and have been given a response of 'This does not violate our community policy.' That's great." (P8)

In some cases the platform's response to reported content seemed like an implicit endorsement of it.

> "I know I've reported things a few times and basically got annoyed because Facebook said it didn't count within their community guidelines - and I'm like, the person said a racial slur. What do you mean it's not within your community guidelines." (P12)

In addition to hopelessness that reporting content would have an effect, our participants were deterred because the labor of reporting all of the harmful content that they are subject to would need to be someone's "full-time job" (P18). This would seem to be a good candidate for algorithmic content moderation [71]. However, similar to their experiences with manual reporting, our participants' past experiences with algorithmic moderation had led them to not trust that algorithms would recognize the intricacies of the types of harm they were subject to or could help protect them:

> "I do not trust an algorithm to stop people from saying they are going to blow me up or something." (P15)

Finally, many of our participants did not necessarily see removal of harmful content as a high priority for themselves, especially because they believed that due to the fast pace of social media many people would have seen the content by the time it was removed anyway. Some even saw value in allowing the harmful content to remain, for instance for educational purposes. Others talked about the importance of preserving harmful content, particularly threats, as evidence in case the situation escalates.

> "I don't delete bad things. [...] People ask me, 'Why don't you?' And I tell them, 'This is the reality. And if I delete them then you get to escape the reality that I have to live. So even though it may be offensive and hurtful, it is reality. And you don't get to deny it just because you don't get to see it. So I'm going to help you see it. That's my job. It even made the news in [anonymized U.S. state]. It ended up - an article came out about it. Not through me. They observed it, or people reported. An article came out and even Senator [anonymized] who was my opponent at the time even tweeted in support of me and that he was sorry that people were acting this way." (P6)

While our participants mostly did not report harmful content, there were exceptions. P8 described a case where they were the target of a misinformation campaign that put them in danger with the Turkish government and were able to get the content removed:

> "I have ninety-nine problems and I don't want the Turkish government as another one. So I reported it to Facebook, and they shockingly said it did violate – wow! – it did violate community policy and they actually took it down. So that was good." (P8)

Our findings illustrate the limitations of content moderation in removing harmful content when the harm is sustained over time and space or is particular to a marginalized identity group. We also found that the mere removal of offending content was not always a priority for our participants. This finding mirrors research from Musgrave et al. who found that Black women distrust social media platforms in reporting instances of harm [116]. Similar to our sample of Muslim Americans, some participants in that study had tried to report offending content with no results. A few had had their own accounts taken down. We will discuss these limitations and potential paths forward in the discussion section.

5.6 Platform affordances that give users control are useful

Our participants described the most useful platform affordances for protecting themselves to be those that afforded them higher control of their audience and who can contact them. Prior research has studied the role that collective block-lists can play in protecting victims of online

harassment [92]. Similarly, P8 described their process for using Facebook's friend feature as a white-list to manage who can communicate with them. Since we conducted this interview, Twitter has implemented a similar feature to limit responses to Tweets.

> "On Facebook I've turned off comments to non-friends, to people I am not friends with, and so everyone who comments is my "friend" and probably doesn't want to get unfriended. Those are the only people that are allowed to comment. So I've learned along the way to manage my mental health [...] As far as I'm concerned it's my page, and I don't need to consume toxic waste. Twitter – there's no way to control that, and so there's a lot of negative [content ...] I've taken a step back from Twitter because it's so, so toxic. It's just a cesspool of white supremacists and Islamophobes." (P8)

Almost all of our participants found blocking a particularly useful affordance. Although, sometimes they struggled to reconcile blocking people with their personal and religious values and positions as public figures:

> "I've become a lot more liberal with the block button, I'd say. That has been something that I've learned over time. There is something a little shameful about blocking someone because part of you feels like you are coinciding, that you can't win this argument, that this person won't change - it's a little un-Islamic - we are supposed to believe everyone can change, you are supposed to be your best version so that people can see it, but you know, that's what I was taught to do Dawah. So, something feels a bit un-Islamic about blocking people, that's why I was hesitant, but nowadays I don't really care. I'm just blocking anyone who shares something that is harassing, that is beyond the pail." (P1)

Overall this finding highlights the potential for designing strategies that afford people more information and control to protect themselves.

6 DISCUSSION

We found that Muslims in the U.S., a heavily marginalized group, have created counterpublics that strive to influence wider dominant publics, but are severely limited due to rampant online harms such as harassment, hate speech, misinformation, and doxing. We argue that the severity and scale of this harm over time and space is not an accident or anomaly, but a coordinated practice of gaining and maintaining spatial and racial control over social media platforms like Twitter [28, 99]. Based on our findings, we focus our discussion on future directions for better supporting counterpublics and marginalized people online.

6.1 Content Moderation is Fundamentally Limited

Our findings confirm prior research arguing that content moderation's impact on people of marginalized identities is to either conform to dominant norms or to resist [57]. Researchers have proposed potential solutions for improving content moderation, the current dominant framework for addressing online harm. Some researchers have proposed jury solutions [53, 166] or advisory boards [95] to address disagreements over what content violates site guidelines. These strategies may not solve disproportionate removal issues for marginalized people who are unlikely to be represented on such panels [79], therefore researchers have proposed using people's identity to form balanced and representative juries [70]. Other researchers have called for more transparency, accountability, and explanations [32, 67, 118] as well as contestability [96] of moderation decision.

While we agree with these calls, our findings and prior work have led us to conclude that content moderation – as almost the sole design and policy intervention deployed by social media platforms to curb harmful content – has not been effective in reducing harms against marginalized people [116]. This is partly because the needs of marginalized people who are harmed is not limited to just

removing the content [162]. In some cases, rather, content removal policies have exposed these group to new forms of harm, such as mass reporting campaigns and account take downs [79]. Therefore, we do not expect that more representation, transparency, and accountability on the same system will be enough to protect counterpublics from rampant online harm that stretches over time and space and weaponizes affordances and identity.

A central limitation of content moderation as the sole mechanism for addressing online harm is that it assumes that harm stems from individual pieces of content, instead of from people and their relationships [135]. To effectively address such harm, we need plans of action for how to move forward when it does inevitably occur [82]. We turn to philosophies and practices of restorative and transformative justice for potential paths forward.

6.2 Restorative Justice and Protection from Harm

Restorative justice is a philosophy and practice of justice that views harm not as a crime against the rules of the state (or the platform), but as a violation of people and their interpersonal relationships [165]. Violations create obligations, and the central obligation of restorative justice is to right the wrong. At minimum, restorative justice requires that we address the victim's needs related to the harm; hold offenders accountable to right those wrongs; and involve victims, offenders, and communities in this process [165]. Recently, HCI researchers have studied what restorative justice approaches to addressing online harm might look like [82, 83, 116, 139, 162]. This work has identified what adolescents need to heal from online harm [162] and gathered preferences from different groups of users for restorative based actions such as apologies [139]. Researchers have also advocated for the design of affordances for mediation and reconciliation [18]. Finally, researchers have considered how the concept of subsidiarity, rather than scalability, might enable large scale implementations of restorative justice practices in online governance [83].

Here, we focus on one central aspect of restorative and transformative justice practices: protecting people from harm [22, 113, 153, 165]. This practice can mean protecting people before harm has happened through community building, education, and making values, rules, and accountability mechanisms visible. It can also mean protecting people who have been harmed from future harm. We discuss what protecting people from the online harms that we identified can look like in practice.

We found that Muslim Americans are harmed in ways that extend over time and space. This made it extremely costly for them to report content that targeted them. It is important to note that such mass harassment is not naturally occuring, but is made possible because of specific affordances of social media platforms. Affordances greatly shape how we connect and interact with others [50, 140, 156]. Frictionless sharing of text, images, and videos to large and diverse audiences make harassment easier [158]. One path forward can be for platforms to implement graduated privileges for interacting with other users, particularly ones who the user is not well connected to. This means that new accounts may not have the same access and affordances as other accounts, such as sending images through direct messages. Over time, accounts gain new privileges that may be taken away if they engage in harassment. We also found that when they were under mass attacks, it became impossible for our participants to use their social media accounts for other purposes such as support seeking. One path forward can be a "bunker" mode for social media accounts that limits the user's accessibility for harm but also provides them with the information and social support they need to protect themselves [162], such as how many attempts to contact them were made or if their personal information was shared online. Vitak et al. propose a similar "limited" mode for targets of harassment where only a subset of users can interact with them as an alternative to deactivating one's account [158].

We also found that the people who targeted our participants weaponized platform affordances in particularly harmful ways, such as continuously tagging them on violent images over months.

Addressing these issues will require a form of "stress testing" platform affordances for their potential to be weaponized and designing safety mechanisms to counter weaponization such as tagging or blurring potentially harmful images [79]. This also includes the new affordances that we have proposed above.

In creating just futures we can not however rely solely on localized harm prevention or even on repairing one harm at a time [114]. Our research shows that the scale and severity of the harm that Muslim Americans and other marginalized groups face is much bigger than what individual interventions may effectively address. Transformative justice is a philosophy and practice that seeks to make visible and eradicate the root causes of harm such as white supremacy and hetero-patriarchy. A systemic analysis of the harms that our participants face does not only surface islamophobia as an underlying cause of harm, but also makes visible platform politics and business models as driving forces for the affordances that enable harm. Platforms prioritize maximum engagement and broad content distribution over other values such as user safety. In other words, hate is good business [26]. Platforms also hope that governance mechanisms, such as content moderation, either disappear under the surface or are perceived as rational and fair policing [35]. Restorative and transformative justice approaches will do neither, as the work of repair and healing from harm and transforming its underlying root causes is costly and challenging, yet deeply rewarding. Transformative justice teaches a politics of liberation [10, 113]. Such an approach to addressing online harm will prioritize healing and joy for those who are harmed the most such as Black women [116], will practice collective action and mutual support [44, 136, 153], and will likely require that platforms be reconceptualized entirely to be community owned and governed [30, 77–79].

7 CONCLUSION

This paper examined how visible Muslim Americans are harmed when they engage in counterpublic activities online. We found that the harm that our participants experienced was sustained over long periods of time and space, weaponized platform affordances, and weaponized gender and identity. We found that Muslim Americans used strategies such as building an audience and limiting information about and access to themselves online to be able to use their voice safely. We also found that harm, and strategies for protection were gendered. Finally, we discuss the limitations of content moderation as the dominant framework for addressing online harm as well as present potential paths forward.

ACKNOWLEDGMENTS

We would like to thank our interview participants and Muslim American advocacy groups for sharing their knowledge and experiences with us. We would also like to thank Sarah Sakha for her feedback on the earlier version of this draft. This research was supported by Facebook's content policy research award.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. arXiv:2101.05783 [cs.CL]
[2] Claire L Adida, David D Laitin, and Marie-Anne Valfort. 2010. Identifying barriers to Muslim integration in France. *Proceedings of the National Academy of Sciences* 107, 52 (2010), 22384–22390.
[3] Patricia A Adler and Peter Adler. 1987. *Membership roles in field research*. Vol. 6. Sage.
[4] Sarah A Aghazadeh, Alison Burns, Jun Chu, Hazel Feigenblatt, Elizabeth Laribee, Lucy Maynard, Amy LM Meyers, Jessica L O'Brien, and Leah Rufus. 2018. GamerGate: A case study in online harassment. In *Online harassment*. Springer, 179–207.
[5] Syed Ishtiaque Ahmed. 2022. Situating ethics: a postsecular perspective for HCI. *Interactions* 29, 4 (2022), 84–86.

[6] Ebtisam Alabdulqader, Norah Abokhodair, and Shaimaa Lazem. 2017. Human-computer interaction across the Arab world. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 1356–1359.
[7] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 3906–3918.
[8] Molly Andrews. 2002. Introduction: Counter-narratives and the power to oppose. (2002).
[9] Mariam Asad. 2019. Prefigurative Design as a Method for Research Justice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 200 (Nov. 2019), 18 pages. https://doi.org/10.1145/3359302
[10] Mariam Asad. 2019. Prefigurative design as a method for research justice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–18.
[11] Robert Asen. 2000. Seeking the "counter" in counterpublics. *Communication theory* 10, 4 (2000), 424–446.
[12] Zahra Ashktorab and Jessica Vitak. 2016. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 3895–3905.
[13] Houston A Baker Jr. 1994. Critical memory and the black public sphere. *Public Culture* 7, 1 (1994), 3–33.
[14] Elyas Bakhtiari. 2020. Health effects of Muslim racialization: Evidence from birth outcomes in California before and after September 11, 2001. *SSM-Population Health* 12 (2020), 100703.
[15] Tom Bartlett. 2012. *Hybrid voices and collaborative change: Contextualising positive discourse analysis*. Vol. 4. Routledge.
[16] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social forces* (2019).
[17] Khaled A Beydoun. 2018. *American Islamophobia: Understanding the Roots and Rise of Fear*. Univ of California Press.
[18] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When online harassment is perceived as justified. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[19] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in social virtual reality: Challenges for platform governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
[20] Lindsay Blackwell, Mark Handel, Sarah T Roberts, Amy Bruckman, and Kimberly Voll. 2018. Understanding" bad actors" online. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–7.
[21] Brooke Bosley, Christina N Harrington, Susana M Morris, and Christopher A Le Dantec. 2022. Healing Justice: A Framework for Collective Healing and Well-Being from Systemic Traumas. In *Designing Interactive Systems Conference*. 471–484.
[22] John Braithwaite. 2016. Restorative justice and responsive regulation: The question of evidence. *RegNet Research Paper* 2016/51 (2016).
[23] Axel Bruns and Jean Burgess. 2011. The use of Twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European consortium for political research (ECPR) general conference 2011*. The European Consortium for Political Research (ECPR), 1–9.
[24] Brian Robert Calfano, Nazita Lajevardi, and Melissa R Michelson. 2019. Trumped up challenges: limitations, opportunities, and the future of political research on Muslim Americans. *Politics, Groups, and Identities* 7, 2 (2019), 477–487.
[25] Craig Calhoun. 1992. *Introduction: Habermas and the public sphere*. MIT press.
[26] Southern Poverty Law Center. 2016. Update: 1,094 bias-related incidents in the month following the election. *Hatewatch. Retrieved from https://www. splcenter. org/hatewatch/2016/12/16/update-1094-bias-related-incidents-month-following-election* (2016).
[27] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.
[28] Sharad Chari. 2008. Critical geographies of racial and spatial control. *Geography Compass* 2, 6 (2008), 1907–1921.
[29] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring# GamerGate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th international conference on world wide web companion*. 1285–1290.
[30] Aymar Jean Christian, Faithe Day, Mark Díaz, and Chelsea Peterson-Salahuddin. 2020. Platforming intersectionality: Networked solidarity and the limits of corporate social Media. *Social Media+ Society* 6, 3 (2020), 2056305120933301.
[31] Danielle Keats Citron and Mary Anne Franks. 2014. Criminalizing revenge porn. *Wake Forest L. Rev.* 49 (2014), 345.
[32] Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.* 91 (2011), 1435.
[33] Sabina Civila, Luis M Romero-Rodríguez, and Amparo Civila. 2020. The Demonization of Islam through Social Media: A Case Study of# Stopislam in Instagram. *Publications* 8, 4 (2020), 52.
[34] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

[35] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.
[36] Cristina Criddle. 2020. Transgender users accuse TikTok of censorship.
[37] Karam Dana, Matt A Barreto, and Kassra AR Oskooii. 2011. Mosques as American institutions: Mosque attendance, religiosity and integration into the political system among American Muslims. *Religions* 2, 4 (2011), 504–524.
[38] Karam Dana, Nazita Lajevardi, Kassra AR Oskooii, and Hannah L Walker. 2019. Veiled Politics: Experiences with Discrimination among Muslim Americans. *Politics & Religion* 12, 4 (2019).
[39] Karam Dana, Bryan Wilcox-Archuleta, and Matt Barreto. 2017. The political incorporation of Muslims in the United States: The mobilizing role of religiosity in Islam. *Journal of Race, Ethnicity and Politics* 2, 2 (2017), 170–200.
[40] Jessie Daniels. 2013. Race and racism in Internet studies: A review and critique. *new media & society* 15, 5 (2013), 695–719.
[41] Michael C Dawson. 1994. A black counterpublic?: Economic earthquakes, racial agenda (s), and black politics. *Public Culture* 7, 1 (1994), 195–223.
[42] Valentina Di Stasio, Bram Lancee, Susanne Veit, and Ruta Yemane. 2021. Muslim by default or religious discrimination? Results from a cross-national field experiment on hiring discrimination. *Journal of Ethnic and Migration Studies* 47, 6 (2021), 1305–1326.
[43] Jeff Diamant. 2017. American Muslims are concerned–but also satisfied with their lives. (2017).
[44] Larry Diamond. 2010. Liberation technology. *Journal of Democracy* 21, 3 (2010), 69–83.
[45] M Duggan. 2017. Men, women experience and view online harassment differently. *Pew Research Center* (2017).
[46] Maeve Duggan. 2017. Online harassment 2017. (2017).
[47] Christopher E. Smith. 2020. Organizational Support and the Electoral Prospects of Progressive Congressional Candidates: An Inside View from 2018. *New Political Science* 42, 2 (2020), 218–232.
[48] Stine Eckert and Kalyani Chadha. 2013. Muslim bloggers in Germany: an emerging counterpublic. *Media, Culture & Society* 35, 8 (2013), 926–942. https://doi.org/10.1177/0163443713501930 arXiv:https://doi.org/10.1177/0163443713501930
[49] Caleb Elfenbein. 2016. Suit seeks to limit anti-Muslim speech on Facebook but roots of Islamophobia run far deeper. https://theconversation.com/suit-seeks-to-limit-anti-muslim-speech-on-facebook-but-roots-of-islamophobia-run-far-deeper-159418
[50] Nicole B Ellison and Jessica Vitak. 2015. Social network site affordances and their relationship to social capital processes. *The handbook of the psychology of communication technology* (2015), 203–227.
[51] Sheena Erete, Yolanda A Rankin, and Jakita O Thomas. 2021. I can't breathe: Reflections from Black women in CSCW and HCI. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–23.
[52] John L Esposito. 1999. *The Islamic threat: Myth or reality?* Oxford University Press, USA.
[53] 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*. 1–14.
[54] Rita Felski and Joseph Felski. 1989. *Beyond feminist aesthetics: Feminist literature and social change*. Harvard University Press.
[55] Miriam Fernandez and Harith Alani. 2018. Online misinformation: Challenges and future directions. In *Companion Proceedings of the The Web Conference 2018*. 595–602.
[56] Estelle Ferrarese. 2014. Nancy Fraser and the Theory of Participatory Parity. *New Left Review* 86 (2014), 55–72.
[57] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. 2020. Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.
[58] Center for Countering Digital Hate. 2022. *Failure to Protect: How Tech Giants Fail to Act on User Reports of Antisemitism*. Retrieved July 15, 2025 from https://counterhate.com/wp-content/uploads/2022/05/Failure-to-Protect.pdf
[59] Center for Countering Digital Hate. 2022. *Failure to Protect: Social Media Platforms are Failing to Act on Anti-Muslim Hate*. Retrieved July 15, 2025 from https://counterhate.com/wp-content/uploads/2022/05/Anti-Muslim-Hate-Failure-to-Protect.pdf
[60] Center for Countering Digital Hate. 2022. *Hidden Hate: How Instagram Fails to Act on 9 in 10 Reports of Mysogyny in DMs*. Retrieved July 15, 2025 from https://counterhate.com/wp-content/uploads/2022/05/Final-Hidden-Hate.pdf
[61] Nancy Fraser. 2021. Rethinking the public sphere: A contribution to the critique of actually existing democracy. In *Public Space Reader*. Routledge, 34–41.
[62] Diana Freed, Sam Havron, Emily Tseng, Andrea Gallardo, Rahul Chatterjee, Thomas Ristenpart, and Nicola Dell. 2019. " Is my phone hacked?" Analyzing Clinical Computer Security Interventions with Survivors of Intimate Partner Violence. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[63] Ryan J Gallagher, Elizabeth Stowell, Andrea G Parker, and Brooke Foucault Welles. 2019. Reclaiming stigmatized narratives: The networked disclosure landscape of# MeToo. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[64] Meira Gebel. 2020. Black creators say tiktok still secretly hides their content. *Digital Trends. Retrieved December* 9 (2020), 2021.
[65] Sonia Ghumman and Ann Marie Ryan. 2013. Not welcome here: Discrimination towards women who wear the Muslim headscarf. *Human Relations* 66, 5 (2013), 671–698.
[66] James J Gibson. 1977. The theory of affordances. Perceiving, Acting and Knowing. *Eds. Robert Shaw and John Bransford* (1977).
[67] Tarleton Gillespie. 2017. Governance of and by platforms. *The SAGE handbook of social media* (2017), 254–278.
[68] Tarleton Gillespie. 2018. *Custodians of the Internet.* Yale University Press.
[69] Jennifer M Gómez. 2019. What's the harm? Internalized prejudice and cultural betrayal trauma in ethnic minorities. *American Journal of Orthopsychiatry* 89, 2 (2019), 237.
[70] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.
[71] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.
[72] Nitesh Goyal, Leslie Park, and Lucy Vasserman. 2022. " You have to prove the threat is real": Understanding the needs of Female Journalists and Activists to Document and Report Online Harassment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.
[73] Kishonna L Gray. 2012. Intersecting oppressions and online communities: Examining the experiences of women of color in Xbox Live. *Information, Communication & Society* 15, 3 (2012), 411–428.
[74] Jessica Guynn. 2019. Facebook while black: Users call it getting 'Zucked,'say talking about racism is censored as hate speech. *Usa today* 24 (2019).
[75] Jurgen Habermas. 1991. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society.* MIT press.
[76] Yvonne Yazbeck Haddad. 2007. The post-9/11 hijab as icon. *Sociology of religion* 68, 3 (2007), 253–267.
[77] Oliver L Haimson, Justin Buss, Zu Weinger, Denny L Starks, Dykee Gorrell, and Briar Sweetbriar Baron. 2020. Trans Time: Safety, Privacy, and Content Warnings on a Transgender-Specific Social Media Site. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.
[78] Oliver L Haimson, Avery Dame-Griff, Elias Capello, and Zahari Richter. 2021. Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist media studies* 21, 3 (2021), 345–361.
[79] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.
[80] Debarati Halder. 2014. Book Review of Hate Crimes in Cyberspace. *International Journal of Cyber Criminology* 8, 2 (2014), 172.
[81] Christina N Harrington, Katya Borgos-Rodriguez, and Anne Marie Piper. 2019. Engaging low-income African American older adults in health discussions through community-based design workshops. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–15.
[82] A Hasinoff, AN Gibson, and N Salehi. 2020. The promise of restorative justice in addressing online harm.
[83] Amy A Hasinoff and Nathan Schneider. 2022. From Scalability to Subsidiarity in Addressing Online Harm. *Social Media+ Society* 8, 3 (2022), 20563051221126041.
[84] William Hobbs and Nazita Lajevardi. 2019. Effects of divisive political campaigns on the day-to-day segregation of arab and muslim Americans. *American Political Science Review* 113, 1 (2019), 270–276.
[85] William Hobbs, Nazita Lajevardi, Xinyi Li, and Caleb Lucas. 2021. Group Salience, Inflammatory Rhetoric, and the Persistence of Hate Against Religious Minorities. (2021).
[86] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.
[87] Sarah J Jackson, Moya Bailey, and Brooke Foucault Welles. 2020. *# HashtagActivism: Networks of race and gender justice.* Mit Press.
[88] Sarah J. Jackson and Brooke Foucault Welles. 2015. Hijacking #myNYPD: Social Media Dissent and Networked Counterpublics. *Journal of Communication* 65, 6 (2015), 932–952. https://doi.org/10.1111/jcom.12185 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcom.12185
[89] Tracy Jan and Elizabeth Dwoskin. 2017. A white man called her kids the n-word. Facebook stopped her from sharing it. *Washington Post (July 2017). https://www. washingtonpost. com/business/economy/for-facebook-erasing-hatespeech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story. html* (2017).
[90] Ashley Jardina and LaFleur Stephens-Dougan. 2021. The electoral consequences of anti-Muslim prejudice. *Electoral Studies* 72 (2021), 102364.

[91] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.
[92] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.
[93] Kishi Katayoun. 2017. Assaults against Muslims in US Surpass 2001 Level. *Pew Research Center, November* 15 (2017), 2017.
[94] Shamika Klassen, Sara Kingsley, Kalyn McCall, Joy Weinberg, and Casey Fiesler. 2021. More than a Modern Day Green Book: Exploring the Online Community of Black Twitter. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.
[95] Kate Klonick. 2019. The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression. *Yale LJ* 129 (2019), 2418.
[96] Daniel N Kluttz, Nitin Kohli, and Deirdre K Mulligan. 2022. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In *Ethics of Data and Analytics*. Auerbach Publications, 420–428.
[97] Beth Kolko, Lisa Nakamura, and Gilbert Rodman. 2013. *Race in cyberspace*. Routledge.
[98] Rachel Kuo. 2018. Racial justice activist hashtags: Counterpublics and discourse circulation. *New Media & Society* 20, 2 (2018), 495–514. https://doi.org/10.1177/1461444816663485 arXiv:https://doi.org/10.1177/1461444816663485
[99] Rachel Kuo. 2018. Racial justice activist hashtags: Counterpublics and discourse circulation. *New Media & Society* 20, 2 (2018), 495–514.
[100] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3739–3748.
[101] Nazita Lajevardi. 2020. *Outsiders at home: The politics of American Islamophobia*. Cambridge University Press.
[102] Nazita Lajevardi and Marisa Abrajano. 2019. How negative sentiment toward Muslim Americans predicts support for Trump in the 2016 Presidential Election. *The Journal of Politics* 81, 1 (2019), 296–302.
[103] Nazita Lajevardi and Kassra AR Oskooii. 2018. Old-fashioned racism, contemporary islamophobia, and the isolation of Muslim Americans in the age of Trump. *Journal of Race, Ethnicity and Politics* 3, 1 (2018), 112–152.
[104] Nazita Lajevardi, Kassra AR Oskooii, and Hannah Walker. 2022. Hate, amplified? Social media news consumption and support for anti-Muslim policies. *Journal of Public Policy* (2022), 1–28.
[105] Nazita Lajevardi, Kassra AR Oskooii, Hannah L Walker, and Aubrey L Westfall. 2020. The Paradox between integration and perceived discrimination among American Muslims. *Political Psychology* 41, 3 (2020), 587–606.
[106] Nazita Lajevardi and Liesel Spangler. 2022. Evaluating Muslim American Representation. *PS: Political Science & Politics* 55, 2 (2022), 285–290.
[107] Sonia Livingstone, Lucyna Kirwil, Cristina Ponte, and Elisabeth Staksrud. 2014. In their own words: What bothers children online? *European Journal of Communication* 29, 3 (2014), 271–288.
[108] Jessica H Lu and Catherine Knight Steele. 2019. 'Joy is resistance': cross-platform resilience and (re) invention of Black oral culture online. *Information, Communication & Society* 22, 6 (2019), 823–837.
[109] Yusr Mahmud and Viren Swami. 2010. The influence of the hijab (Islamic head-cover) on perceptions of women's attractiveness and intelligence. *Body image* 7, 1 (2010), 90–93.
[110] Adrienne Massanari. 2017. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New media & society* 19, 3 (2017), 329–346.
[111] Sharan B Merriam and Robin S Grenier. 2019. *Qualitative research in practice: Examples for discussion and analysis*. John Wiley & Sons.
[112] Robert Meyer and Michel Cukier. 2006. Assessing the attack threat due to IRC channels. In *International Conference on Dependable Systems and Networks (DSN'06)*. IEEE, 467–472.
[113] M Mingus. 2019. Transformative justice: A brief description. Leaving Evidence.
[114] Mia Mingus. 2022. Transformative justice: A brief description. *Fellowship* 84, 2 (2022), 17–19.
[115] Besheer Mohamed and Jeff Diamant. 2020. Black Muslims account for a fifth of all U.S. Muslims, and about half are converts to Islam. https://www.pewresearch.org/fact-tank/2019/01/17/black-muslims-account-for-a-fifth-of-all-u-s-muslims-and-about-half-are-converts-to-islam/
[116] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. 2022. Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media. In *CHI Conference on Human Factors in Computing Systems*. 1–17.
[117] Maryam Mustafa, Shaimaa Lazem, Ebtisam Alabdulqader, Kentaro Toyama, Sharifa Sultana, Samia Ibtasam, Richard Anderson, and Syed Ishtiaque Ahmed. 2020. IslamicHCI: Designing with and within Muslim Populations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[118] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
[119] Angela Nagle. 2017. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right.* John Hunt Publishing.
[120] Lisa Nakamura. 2013. *Cybertypes: Race, ethnicity, and Identity on the Internet.* Routledge.
[121] Fayika Farhat Nova, Md Rashidujjaman Rifat, Pratyasha Saha, Syed Ishtiaque Ahmed, and Shion Guha. 2018. Silenced voices: Understanding sexual harassment on anonymous social media among Bangladeshi people. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 209–212.
[122] Kassra AR Oskooii. 2016. How discrimination impacts sociopolitical behavior: A multidimensional perspective. *Political Psychology* 37, 5 (2016), 613–640.
[123] Kassra AR Oskooii, Karam Dana, and Matthew A Barreto. 2019. Beyond generalized ethnocentrism: Islam-specific beliefs and prejudice toward Muslim Americans. *Politics, Groups, and Identities* (2019), 1–28.
[124] Roya Pakzad and Niloufar Salehi. 2019. Anti-Muslim Americans: Computational Propaganda in the United States. (2019). https://www.iftf.org/fileadmin/user_upload/downloads/ourwork/IFTF_Anti-Muslim_comp.prop_W_05.07.19.pdf
[125] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*. 369–374.
[126] Lawrence Pintak, Jonathan Albright, Brian J Bowe, and Shaheen Pasha. 2019. Islamophobia: Stoking Fear and Prejudice in the 2018 Midterms. *Social Science Research Council. https://www. ssrc. org/publications/view/islamophobia-stoking-fear-and-prejudice-in-the-2018-midterms* (2019).
[127] Bailey Poland. 2016. *Haters: Harassment, abuse, and violence online.* U of Nebraska Press.
[128] Elissa M Redmiles, Jessica Bodford, and Lindsay Blackwell. 2019. "I just want to feel safe": A Diary Study of Safety Perceptions on Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 405–416.
[129] Mohammad Rashidujjaman Rifat, Firaz Ahmed Peer, Hawra Rabaan, Nusrat Jahan Mim, Maryam Mustafa, Kentaro Toyama, Robert B Markum, Elizabeth Buie, Jessica Hammer, Sharifa Sultana, et al. 2022. Integrating Religion, Faith, and Spirituality in HCI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.
[130] Mohammad Rashidujjaman Rifat, Hasan Mahmud Prottoy, and Syed Ishtiaque Ahmed. 2022. Putting the Waz on Social Media: Infrastructuring Online Islamic Counterpublic through Digital Sermons in Bangladesh. In *CHI Conference on Human Factors in Computing Systems*. 1–19.
[131] Alyssa E Rippy and Elana Newman. 2006. Perceived religious discrimination and its relationship to anxiety and paranoia among Muslim Americans. *Journal of Muslim Mental Health* 1, 1 (2006), 5–20.
[132] Bruce Robbins. 1993. *The phantom public sphere.* University of Minnesota Press.
[133] Sarah T Roberts. 2019. *Behind the screen.* Yale University Press.
[134] Edward Said. 1979. Orientalism. 1978. *New York: Vintage* 199 (1979).
[135] Niloufar Salehi. 2020. *Do No Harm.*
[136] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, and Kristy Milland. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1621–1630.
[137] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-computer Interaction* 2, CSCW (2018), 1–27.
[138] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5412–5427. https://doi.org/10.1145/3025453.3025766
[139] 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.
[140] Andrew Richard Schrock. 2015. Communicative affordances of mobile media: Portability, availability, locatability, and multimediality. *International Journal of Communication* 9 (2015), 18.
[141] Nura A Sediqe. 2020. Stigma Consciousness and American Identity: The Case of Muslims in the United States. *PS: Political Science & Politics* 53, 4 (2020), 674–678.
[142] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.
[143] Saher Selod. 2015. Citizenship denied: The racialization of Muslim American men and women post-9/11. *Critical Sociology* 41, 1 (2015), 77–95.

[144] Saher Selod and David G Embrick. 2013. Racialization and Muslims: Situating the Muslim experience in race scholarship. *Sociology Compass* 7, 8 (2013), 644–655.
[145] Yonat Shimron. 2016. When wearing a hijab becomes too dangerous. *USA Today* 12 (2016).
[146] Jennifer L Simpson and Kimberly Carter. 2008. Muslim women's experiences with health care providers in a rural area of the United States. *Journal of Transcultural Nursing* 19, 1 (2008), 16–23.
[147] Vivek K Singh, Marie L Radford, Qianjia Huang, and Susan Furrer. 2017. " They basically like destroyed the school one day" On Newer App Features and Cyberbullying in Schools. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1210–1216.
[148] Catherine R Squires. 2002. Rethinking the black public sphere: An alternative vocabulary for multiple public spheres. *Communication theory* 12, 4 (2002), 446–468.
[149] Catherine Knight Steele. 2021. Digital black feminism. In *Digital Black Feminism*. New York University Press.
[150] Evan Stewart, Nazita Lajevardi, Roy Whitaker, and Tarah Williams. 2022. Who are LGBT Muslims? Challenging Myths and Misnomers. *PRRI Spotlight Analysis* (2022).
[151] Alexis Straka. 2020. *Muslim Americans & Electoral Democracy in the Trump Era.* Ph. D. Dissertation. University of Cincinnati.
[152] Sharifa Sultana, Mitrasree Deb, Ananya Bhattacharjee, Shaid Hasan, SM Raihanul Alam, Trishna Chakraborty, Prianka Roy, Samira Fairuz Ahmed, Aparna Moitra, M Ashraful Amin, et al. 2021. 'Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
[153] Sharifa Sultana, Sadia Tasnuva Pritha, Rahnuma Tasnim, Anik Das, Rokeya Akter, Shaid Hasan, SM Raihanul Alam, Muhammad Ashad Kabir, and Syed Ishtiaque Ahmed. 2022. 'shishushurokkha': A transformative justice approach for combating child sexual abuse in bangladesh. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.
[154] Sean Tackett, J Hunter Young, Shannon Putman, Charles Wiener, Katherine Deruggiero, and Jamil D Bayram. 2018. Barriers to healthcare among Muslim women: a narrative review of the literature. In *Women's Studies International Forum*, Vol. 69. Elsevier, 190–194.
[155] Alexandra To, Wenxia Sweeney, Jessica Hammer, and Geoff Kaufman. 2020. " They Just Don't Get It": Towards Social Technologies for Coping with Interpersonal Racism. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–29.
[156] Jeffrey W Treem and Paul M Leonardi. 2013. Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association* 36, 1 (2013), 143–189.
[157] Jonathan Vanian. 2017. Twitter Toughens Rules on Nudity and Revenge Porn| Fortune. *URL: http://fortune.com/2017/10/27/nudity-revenge-porntwitter* (2017).
[158] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1231–1245.
[159] Jessica Vitak and Jinyoung Kim. 2014. " You can't block people offline" examining how facebook's affordances shape the disclosure process. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 461–474.
[160] Emily A Vogels. 2021. The state of online harassment. *Pew Research Center* 13 (2021).
[161] DE Wittkower. 2017. Disaffordances and dysaffordances in code. *AoIR Selected Papers of Internet Research* (2017).
[162] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. In *CHI Conference on Human Factors in Computing Systems*. 1–15.
[163] Sijia Xiao, Danaë Metaxa, Joon Sung Park, Karrie Karahalios, and Niloufar Salehi. 2020. Random, messy, funny, raw: Finstas as intimate reconfigurations of social media. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[164] Jillian C York. 2022. *Silicon values: The future of free speech under surveillance capitalism.* Verso Books.
[165] Howard Zehr. 2015. *The little book of restorative justice: Revised and updated.* Simon and Schuster.
[166] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: building governance in online communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 365–378.

Received January 2022; revised July 2022; accepted November 2022