"Ignorance is not Bliss": Designing Personalized Moderation to Address Ableist Hate on Social Media

Sharon Heung
Information Science
Cornell Tech
New York, New York, USA
ssh247@cornell.edu

Lucy Jiang
Human Centered Design and Engineering
University of Washington
Seattle, Washington, USA
lucjia@uw.edu

Shiri Azenkot
Information Science
Cornell Tech
New York, New York, USA
shiri.azenkot@cornell.edu

Aditya Vashistha
Cornell University
Ithaca, New York, USA
adityav@cornell.edu

Abstract

Disabled people on social media often experience ableist hate and microaggressions. Prior work has shown that platform moderation often fails to remove ableist hate, leaving disabled users exposed to harmful content. This paper examines how personalized moderation can safeguard users from viewing ableist comments. During interviews and focus groups with 23 disabled social media users, we presented design probes to elicit perceptions on configuring their filters of ableist speech (e.g., intensity of ableism and types of ableism) and customizing the presentation of the ableist speech to mitigate the harm (e.g., AI rephrasing the comment and content warnings). We found that participants preferred configuring their filters through types of ableist speech and favored content warnings. We surface participants' distrust in AI-based moderation, skepticism in AI's accuracy, and varied tolerances in viewing ableist hate. Finally, we share design recommendations to support users' agency, mitigate harm from hate, and promote safety.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; *User studies*; Empirical studies in accessibility; • **Social and professional topics → People with disabilities**.

Keywords

personal content moderation, word filters, platform governance, ableism, hate and harassment

ACM Reference Format:
Sharon Heung, Lucy Jiang, Shiri Azenkot, and Aditya Vashistha. 2025. "Ignorance is not Bliss": Designing Personalized Moderation to Address Ableist Hate on Social Media. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713997

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713997

1 Introduction

Disabled[1] people experience high levels of harassment online, including ableist[2] microaggressions [43] and hate [26, 42, 59, 65, 74, 79]. Research has shown that experiencing ableism online can have lasting effects on disabled users' well-being, influencing behaviors such as self-censorship [42] on social media. Furthermore, platform moderation often fails to effectively manage disability-related content, allowing some harmful ableist content to remain while mistakenly removing other legitimate content related to disability [42, 59, 73, 76]. This disconnect between platform moderation and user needs highlights the potential for alternative moderation techniques to help users avoid exposure to ableist hate and harassment.

Personal moderation allows users to configure or customize some aspects of their moderation preferences, based on the content posted by other users [47]. This configuration only affects the user's view, meaning the content remains visible to other users. Currently, platforms offer some personal content moderation tools (e.g., word filters [45], toggles [72], and sensitivity sliders [1]) that enable users to specify types of content they wish to avoid. Platforms also provide account-based moderation tools, letting users choose which accounts to follow and block. In addition to these platform-provided tools, several third-party applications have emerged, empowering users to personalize their content. For example, Google's Tune Chrome Extension [9] allows users to select levels of toxicity, and Bodyguard [5] promises that their AI system *"instantly removes toxic, spam, and damaging content."* The emerging prevalence and promise of these tools suggests that AI-based personal content moderation has the potential to help users avoid harmful and triggering content.

While personalized moderation has potential to reduce the harms of seeing hateful content, few scholars have conducted empirical studies to gather users' perceptions on the usability and efficacy of such tools. For example, Jhaver et al. [47] gathered social media users' perceptions on personal moderation tools based on toxicity, identifying critical needs such as clearer definitions of toxicity, more granular control options, and more transparency through examples

[1] We use identity-first language because it aligns with the language the majority of our participants used.

[2] *Ableist* is the adjective form of the noun *ableism*, used to describe discrimination and prejudice towards disabled people [23, 30]

of filtered content. We extend prior work by understanding how these tools should be designed for identity-based hate — specifically, ableism.

Personalized moderation systems that particularly address identity-based harms have started to emerge. For example, Intel's Bleep [72] filters harmful content related to ableism, racism, and sexism. However, little is known about how effective these identity-based personal moderation tools are, whether these tools account for users' specific needs, and how these tools might be designed to better support those who may benefit from them. To our knowledge, there is no understanding of how disabled people perceive these filters that are designed to address ableism.

To address this critical gap, we examined how an ableism-specific personal moderation tool should be designed to account for the needs and preferences of disabled people. More specifically, we ask:

RQ1: How do disabled social media users perceive and use existing personal moderation tools for addressing ableist hate and harassment?

RQ2: How can personal moderation tools be designed to account for ableist speech during experiences of ableist hate and harassment?

We conducted a two-part study. First, we held initial interviews with 23 disabled social media users to introduce current personal moderation tools and gain insight on their experiences using word filters and blocking. Then, we facilitated eight 90-minute focus groups with 2-3 participants each, presenting design probes to gather feedback on new designs of AI-based personalized moderation tools for addressing ableist content. Our design probes include two ends of a personalized moderation system. First, we present various ableism-specific filter settings (e.g., binary toggle for ableism or toggle for different types of ableist speech). We then show various designs augmenting the presentation of the filtered speech (e.g., having AI rephrase the ableist comment to be less toxic or having a content warning). This contrasts with most existing filters which only fully remove the hateful comment from view. We used these design probes to elicit participants' perceptions on why they would or would not use such tools, the desired capabilities of such tools, and AI's capability of effectively identifying and rephrasing ableist text.

We found that the majority of participants commonly addressed ableist hate and harassment by responding to the perpetrator and/or blocking certain accounts. A subset of participants had experience setting word filters (e.g., removing the r-word); however, word filters were laborious to set up and not always reliable. In response to the design probes, the majority of participants preferred configuring their filters based on types of ableist hate, as it was perceived to be more understandable regarding what types of content would be filtered. Participants also favored content warnings for empowering them to decide whether or not to view the hate. Participants also expressed concerns that AI filters would struggle to identify ableist speech, leading to the wrongful filtering of disability-related content (e.g., posts containing reclaimed words). Since personal moderation only affects an individual user's view, participants wanted platform moderation to also adopt ableism-specific content warnings as a means of educating other users on ableism.

Based on our findings, we discuss how personal moderation can better support the safety of users experiencing ableist hate and harassment online. We recommend for ableism-specific AI filters to move away from full removal of hateful content and instead allow for customized designs (e.g., content warnings) and nudges (e.g., notification of death threats) to support user safety and to promote user agency. We also recommend personal moderation tools to allow users to customize their filters based on specific types of ableist hate. Given the widespread distrust in platform moderation, we recommend increasing transparency and incorporating mechanisms for user control, ensuring that users can oversee and reverse filtering decisions as needed to build trust. In summary, our study contributes:

- Insights into how disabled social media users perceive and utilize existing personal moderation tools (RQ1).
- Designs of AI-based moderation and perceptions of disabled people on AI's capabilities of identifying ableist text (RQ2).
- Design recommendations for how personal moderation can be designed to address ableist speech and support user agency and safety during experiences of online hate (RQ2).

2 Related Work

Like other "isms" (e.g., racism and sexism), ableism is discrimination towards a social group, specifically disabled people [23, 30]. Disability studies scholars have investigated how ableism surfaces societal perceptions of disability. For example, Campbell describes ableism to cast disability as a diminished state of being human [18], which leads to the compulsory preference for non disability [19]. Scholars have also understood how ableism is connected to ideals and attributes that are valued or not valued [95]. It is ableist to assert preference for a child to read print rather than Braille or to walk rather than use a wheelchair, which is harmful for students receiving disability accommodations in school [41]. More broadly, scholars have described ableism as a capitalistic ideology of assigning value to people's productivity. Talila A. Lewis, a disability activist and lawyer, defines ableism as a "system of assigning value to people's bodies and minds based on societally constructed ideas of normalcy, productivity, desirability, intelligence, excellence, and fitness" [57]. We draw on prior work, using ableism as a term to address the collective discriminatory experience the disability community face *online* with a specific focus on *ableist text.*

In this section, we first review HCI literature on ableist speech within social media. Then, we situate our work within the broader context of online moderation and personal moderation, especially with regards to the experiences of people with marginalized identities.

2.1 Ableist Speech on Social Media

Prior work has categorized the diverse ways in which the disability community encounters ableist hate and harassment online, through public comments to private messages[15, 26, 42, 54, 59, 74, 79]. Sannon et al. [79] found that disability activists often experienced invalidating comments on their disability, sexual harassment, fetishization, and coordinated attacks in response to their advocacy work. Building on this work, Heung et al. [42] developed a taxonomy of ableist hate found across 50 disabled content creators with varying disability identities. This taxonomy includes five broader categories of ableist hate (i.e., Slurs & Derogatory language, Violent & Eugenics-related speech, Questioning Ability & Denying Access, Mocking & Invalidating Disability Identity, Objectifying the Disabled Body) and 11 specific types of ableist hate (e.g., Short Slurs; Using Disability as an Insult; Death Threats, Suicide, and Self-harm). The researchers also explored how creators' intersectional identities (e.g., race and sexuality) impacted the frequency of ableist hate, finding that LGBTQ creators experience significantly more ableist hate than non-LGBTQ creators. Heung et al. also alluded to potential differences in ways ableist hate is experienced depending on one's disability identity. For example, people with invisible disabilities (less visibly apparent disability) acknowledged that they may be less likely to experience overt forms of ableist hate, but more often experienced invalidating comments about their disability. Prior work has also shown that invalidation of disability is prominent in online ADHD communities [26].

In addition to overt hate, disabled people also experience microaggressions, or subtle forms of ableist speech. For example, many receive patronizing comments (e.g., "you're so inspirational") and infantilizing remarks (e.g., "where's your mom?") [43, 55]. Prior work acknowledges that the distinction between microaggressions and overt hate is not always clear-cut; for instance, accusations of faking one's disability were perceived to be both a microaggression and an act of overt ableist hate [42, 43].

Researchers have also examined the harms of microaggressions and overt hate and found that these experiences have significant impact on disabled people, including increased emotional distress, anxiety posting online, and self-censorship overtime [42, 43, 52, 79]. Although viewing hate can be harmful, previous research indicates that some users, particularly content creators, may tolerate certain forms of hate [42, 54, 79, 89]. For disability activists and creators, exposure to hate can inspire their activism and inform their educational content in creative ways [42, 79]. For example, Duval et al. [25] found TikTok videos advocating for disabled people or debunking disability stereotypes as a common form of playful content (e.g., using upbeat sound effects and humor). Also on Tiktok, Wang et al. [93] found autistic creators to create hashtags directly in response to ableism, such as #ableismisntcute and #ableistsuck.

We build on existing literature understanding ableism online by exploring how personal moderation can be designed to reduce the harm caused by exposure to ableist text. Our work is motivated by the failure ofplatform moderation in supporting disabled social media users during online hate, which we discuss next.

2.2 Platform Moderation

Content moderation is the organized practice of screening and controlling for unwanted content, content that is deemed as "irrelevant, obscene or illegal" [28, 83]. Most common conceptions of moderation are of mechanisms deployed after an infraction occurs, also known as reactive moderation [35]. Current reactive moderation techniques include filtering or removing inappropriate content, suspending the offending users, or even recommending and curating alternate content [32, 63]. Moderation can be done by human moderators, users themselves, or, increasingly via algorithms running AI models and toxicity classifiers. AI-based reactive moderation can take multiple forms, including automatically filtering out keywords, such as the AutoModerator bot [3] on Reddit. Other techniques leverage natural language processing techniques to automatically detect toxicity (e.g., Perspective API [2], AWS [96]) or computer vision to detect violent or graphic content [68]). While these techniques do not necessarily remove the content from the platform, they often hide the detected content behind a warning — for example, Meta's current warning reads, *"Sensitive content: this video contains content that some people may find upsetting"* [6].

However, despite AI's ability to moderate at scale [33], there are several pitfalls. AI-powered moderation systems often lack context and community-specific nuance, which is especially important since online discourse varies greatly depending on the audience, the place of communication, the speaker, and their tone. For example, Oliva et al. [70] found that drag queen Twitter accounts were considered to have higher perceived levels of toxicity than Donald Trump and white nationalists when moderated by AI models unfamiliar with their lexicon. In addition, AI-based moderation systems are known to exhibit ableist, sexist, colonialist, and racist tendencies. For example, Shahid et al. [86] show that Meta's AI-based moderation has high false positive rate for users in the Global South and the underlying algorithms imbibe coloniality by centering Western norms and erasing minoritized expressions. While these are a few examples of AI's shortcomings, this demonstrates the risk of further silencing already marginalized voices.

Particularly for disabled people, research shows that platform moderation is often inadequate in protecting them from ableist hate and harassment. For example, disabled creators perceived that ableist hate is oftentimes not removed by the platform, despite being reported, forcing creators to manually delete hateful comments themselves or organize other users to report on their behalf [42, 79]. Furthermore, disabled social media users have felt wronged by moderation, such as being penalized by moderation when responding to trolls [59] or by social media algorithms suppressing disability-related content [20, 42, 53, 59, 73]. Platforms not addressing ableist hate may contribute a chilling effect for the disability community, with fear of being overshadowed by hostile voices [11]. Furthermore, with the ableist hate not removed, other disabled users may be less likely to post online; the Pew Research Center reported that 27% of Americans have refrained from posting online after witnessing harassment [22]. Beyond platform moderation not removing ableist hate, disabled users also exert additional labor given the inaccessibility of social media platforms more generally [59, 69].

In our work, we explore how existing and imagined moderation techniques, including personalized moderation, are potentially suited to prevent viewing ableist hate and harassment in particular.

2.3 Personalized Moderation

Given varying norms across cultures and communities [50, 84], research shows that a one-size-fits-all approach to content moderation is insufficient to meet the diverse needs of users [21]. Additionally, platforms have varying definitions of harassment and different corresponding platform policies, highlighting inconsistencies on defining and moderating online harassment across platforms [71]. To maintain free speech online while mitigating harmful content, there is a growing call to move away from a centralized moderation to a user-centered approach [21, 27, 84]. Essentially, what if users themselves could decide how they want their content to be moderated?

Jhaver et al. [47] define personal moderation as tools that *"let users configure their preferences for the activity they want to avoid."* It is important to note that this form of moderation only changes the configured user's view, other social media users can still view the filtered content. Fukuyama et al. [29] refers to this individualized approach to customization of content moderation as *"middleware,"* imagining third-party services as adding an editorial layer between platforms and users. In this section, we describe the two types of personalized moderation, account-based moderation and content-based moderation, specifically highlighting its usage in the context of hateful content.

2.3.1 Personal Account Moderation. Personal account moderation tools enable users to mute or block a particular account, determining who they want to engage with online. Blocking or muting accounts means that the content from that account or creator will no longer appear on a user's feed [7, 10]. Blocking is more restrictive than muting: a user can still interact with and view content from a muted account if they are on their profile, whereas blocking disallows a user from engaging with the other user in any way and is typically known to both parties. Blocking is typically on an individual basis; however, blocklists have emerged as a way for users to easily block many users at once and have been found to be effective in addressing online harassment [46]. Prior work has shown that disabled content creators leverage blocking to foster a safe space for themselves and their followers [42].

2.3.2 Personal Content Moderation. Personal content moderation allows users to make moderation decisions on individual posts based on their content alone, regardless of its source [47]. Common personal content moderation tools include word filters or the ability to mute specific keywords [45]. Prior work has found word filters useful for automatically removing toxic comments and removing potential doxxing attempts, which is the non-consensual release of private and personal information [78, 89].

Word filters on most platforms use rule-based automation, which can require labor of inputting words and variations of words on their own. In response to this difficulty, researchers have developed FilterBuddy [45], which allows creators to easily edit filters, including adding spelling variants of keywords, previewing the effects of specific word filters, importing word filter categories (e.g., Homophobia, Pejorative Terms for Women, and Anti-Black Racism), and sharing word filters with other creators. This type of individual rule-setting is a type of distributed content moderation [49] where content creators have governance over enforcing local rules in the comments. Filtering keywords empowers content creators to automatically moderate their own account, and it is the only tool that enables end-users to preemptively limit harmful content on their profile. One can think of this as a reactive approach, hiding comments after it is posted.

Beyond rule-based filters, personal content moderation tools have begun to integrate AI to identify and therefore filter certain types of content. Some platforms have incorporated these AI-based personal moderation tools. For instance, Twitch creators have Automod, an automated moderation tool that allows users to filter content based on these categories: discrimination and slurs, sexual content hostility, and profanity [8]. Instagram has incorporated sensitivity sliders [1] that default to "normal," but users can choose "more" or "less" to indicate the amount of sensitive content users want to see in their timeline. Similarly, Google released an experimental Chrome browser extension, Tune, that allows users to customize how much toxicity they wish to see in comments across the internet [51].

Despite growing interest in AI-based personal moderation tools, empirical work on end users' perceptions on such tools is limited. One survey study found users to view personal moderation tools as a means for greater agency over their social media experience, and not an infringement on free speech [48]. In another study, Jhaver et al. [47] investigated end users' perceptions on personal content moderation tools that filtered content based on toxicity and identified several improvements to content moderation tools, such as increased clarity in definitions of what is hateful, more granularity in end user controls, and greater transparency in what content gets filtered.

In this paper, we extend work on personalized moderation by capturing disabled social media users' experiences with personal moderation tools (RQ1) and enriching knowledge on how these tools should be designed to account for ableist speech during hate and harassment (RQ2).

3 Methods

We conducted 23 interviews and eight focus groups to understand participants' prior moderation experiences and ideate on personalized moderation techniques.

3.1 Participants

Given that our work focuses on personalized moderation during and after experiences of ableist hate, we specifically selected participants who are social media users and havedisclosed their disability identity on social media. This approach, used by other scholars [42, 43] ensured that participants had firsthand experience with ableist hate directed at them because of their disability identity. Participants were excluded if they were not comfortable communicating in English or American Sign Language (ASL).

We used convenience sampling from a pool of prior participants who had previously agreed to be contacted for future studies, curated by the first author. We then conducted snowball sampling to recruit additional participants. We launched a short screening survey to confirm eligibility. All participants self-identified as having a disability and were 18+ years old. We asked participants to share their disability identity through Blaser et al.'s [14] survey question design with options to select multiple disability identities and input an open-ended response. We recruited a diverse group of participants with varying disabilities and different types of social media usage; 14 of our participants self-identified as content creators and / or influencers. All participants were located in the US, Canada, or Europe. To protect the anonymity of creators, we share aggregated participant demographics in Table 1.

[3] Disability demographics were self-identified and not mutually exclusive, as many participants identified as having more than one disability identity.

Table 1: Aggregated Participant Demographics

Participant Demographics	
Age	18-24 = 3 25-34 = 13 35-44 = 3 45-54 = 4

Gender	Male = 11
	Female = 11
	Trans Male = 1

Race	White = 8 Black & African American = 10 Latina = 1 Latin American = 1 Mixed race (e.g. Hispanic & Asian) = 3
Disability [3]	Blind or low vision = 4 d/Deaf or hard of hearing = 3 Neurodivergent = 4 ADHD = 4 Autism = 7 Health-related disability = 10 Permanent / long-term disability = 10 Physical disability = 1

Social Media Platform	X (previously Twitter) = 20
	Facebook = 19
	Instagram = 19
	TikTok = 12
	Linkedln = 10
	Snapchat = 9
	Reddit = 4
	Twitch = 3
	OnlyFans = 1
	BlueSky = 1

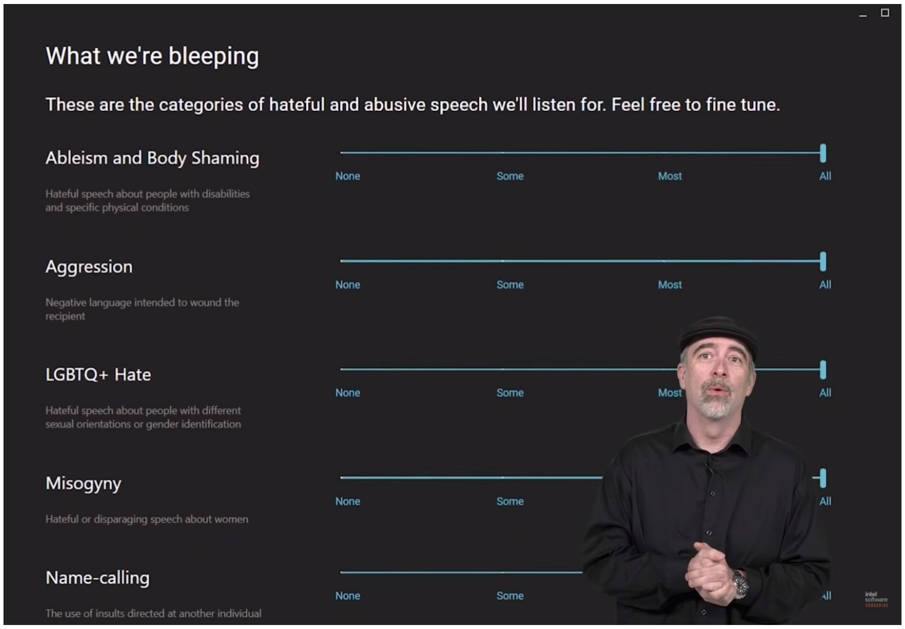
Experiences with Types of Ableist Hate Heung et al. [42]	Short Slurs = 16 Using Disability as an Insult = 19 Death Threats, Suicide, and Self-harm = 8 Violent & Dehumanizing Speech = 14 Eugenics-Related = 9 Disability as Inability = 18 Denial and Stigmatization of Accessibility = 16 Mocking Disability = 17 Accusing of Faking Disability = 13 Attacking Physical Appearance = 13 Sexual Harassment & Fetishization = 9

3.2 Procedure

Eligible participants were invited to participate in (1) a 15-minute interview and (2) a 90-minute focus group, both on Zoom. Participants were compensated with a $10 digital gift card for the interview and a $50 gift card for the focus group. ASL interpreters also consented to participate and were compensated for their time.

During the interviews, we asked participants about their experiences with responding to ableist hate and harassment, introduced existing personalized moderation tools, and gained insight on their experiences and challenges with such tools. Then, we discussed scheduling logistics and accessibility accommodations for the follow-up focus group study. Example of accommodation requests included a verbal description of the design probes during the focus group for blind and low vision participants and flexibility of taking breaks and turning off their video for participants with chronic conditions.

We recognize that discussing ableism may cause emotional distress. Following best practices of conducting research on online hate [81] and a trauma-informed research approach [44], we used the initial interviews to build rapport, reminded participants they could remove themselves from the study or take breaks at any point in time, and shared resources with them to cope with online hate.



What we're bleeping

These are the categories of hateful and abusive speech we'll listen for. Feel free to fine tune.

Ableism and Body Shaming — None Some Most All

Hateful speech about people with disabilities and specific physical conditions

Aggression — None Some Most All

Negative language intended to wound the recipient

LGBTQ+ Hate — None Some Most All

Hateful speech about people with different sexual orientations or gender identification

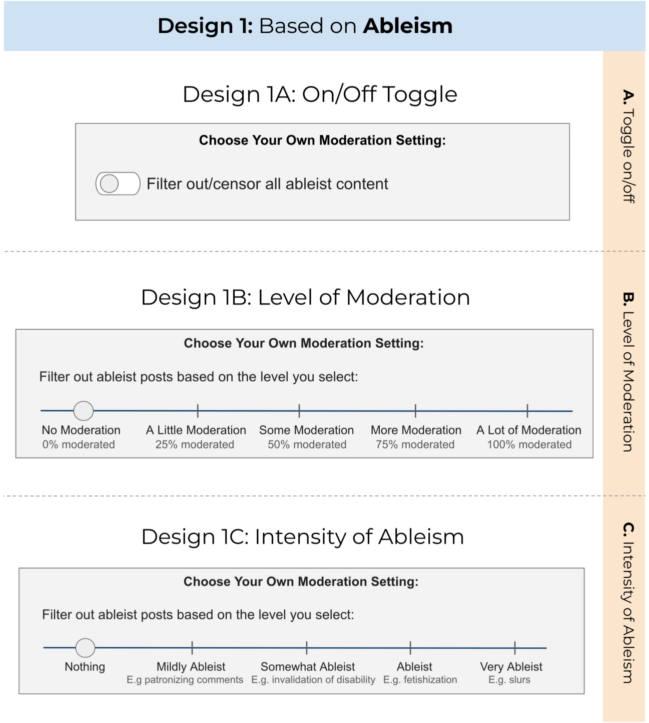
Misogyny — None Some All

Hateful or disparaging speech about women

Name-calling — None Some All

The use of insults directed at another individual

Figure 1: Screenshot of Intel's Bleep interface.



Design 1: Based on **Ableism**

Design 1A: On/Off Toggle

Choose Your Own Moderation Setting:

○ Filter out/censor all ableist content

Design 1B: Level of Moderation

Choose Your Own Moderation Setting:

Filter out ableist posts based on the level you select:

● No Moderation (0% moderated) — A Little Moderation (25% moderated) — Some Moderation (50% moderated) — More Moderation (75% moderated) — A Lot of Moderation (100% moderated)

Design 1C: Intensity of Ableism

Choose Your Own Moderation Setting:

Filter out ableist posts based on the level you select:

● Nothing — Mildly Ableist (E.g patronizing comments) — Somewhat Ableist (E.g. invalidation of disability) — Ableist (E.g. fetishization) — Very Ableist (E.g. slurs)

Figure 2: Diagram of Design Probe 1, configuring filters based on ableism. This includes 3 different designs: 1A (toggling ableist content), 1B (slider based on quantity of ableist posts), and 1C (slider based on intensity of ableism).

After the interviews, we facilitated eight focus groups, seven of which had three participants and one had two participants. We started the focus group with introductions, expectations for creating a safe space, and preliminary feedback on Intel's Bleep design to configure content based on ableism (see Figure 1). This served as an introduction to AI-based personal moderation tools that exist today. Then we spent an hour presenting design probes, including configuring filters based on ableism (Figure 2), configuring filters based on types of ableist hate (Figure 3), and customizing the presentation of ableist hate (Figure 4). After presenting each design, we asked participants to share their thoughts about what they liked or disliked, concerns they had, and what they would have changed about the design. At the end, we asked participants to reflect on all the probes and to build their own personalized moderation tool.

The focus group setting provided participants with an opportunity to highlight their personal preferences and contrast them with others' thoughts. We emphasized that the goal of the probes was not necessarily to understand which was "better," but to concretely ideate on how a personalized moderation tool could be designed to mitigate the harm of viewing ableist hate. We positioned the design probes as works-in-progress requiring their expert feedback, reducing the power dynamics between researchers and participants [64]. These probes also provided a starting point for participants to engage without needing to disclose personal stories in a group setting [90].

3.2.1 *Design Probes.*

Our work builds on prior research and practice in the area of personalized moderation [36, 47]. For example, Jhaver et al. [47] captured end users perspectives of a personalized moderation to filter out varying levels of toxicity. Additionally, Intel's Bleep is an AI-powered tool designed to filter out identity-based hate in voice calls. While Bleep is already available for consumer use, there is limited understanding of how disabled people

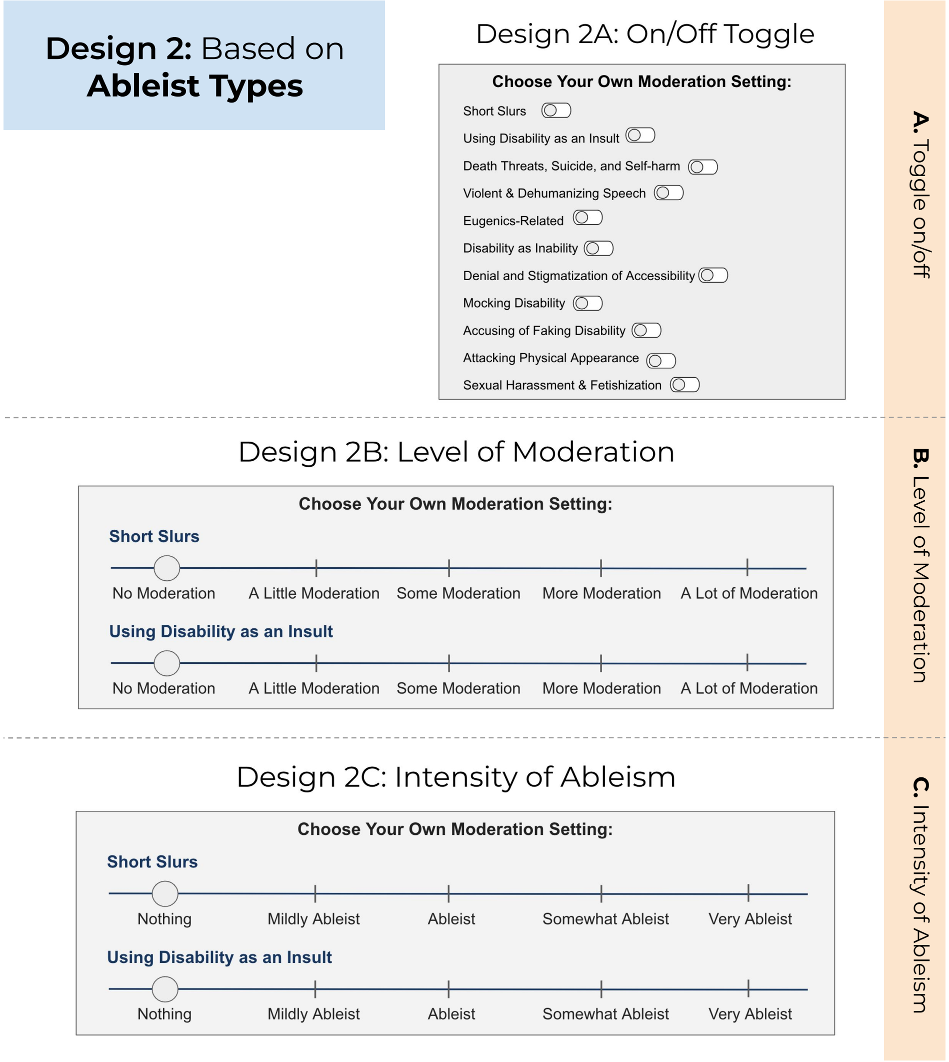

Design 2: Based on Ableist Types
Design 2A: On/Off Toggle
Choose Your Own Moderation Setting:
Short Slurs
Using Disability as an Insult
Death Threats, Suicide, and Self-harm
Violent & Dehumanizing Speech
Eugenics-Related
Disability as Inability
Denial and Stigmatization of Accessibility
Mocking Disability
Accusing of Faking Disability
Attacking Physical Appearance
Sexual Harassment & Fetishization
A. Toggle on/off
Design 2B: Level of Moderation
Choose Your Own Moderation Setting:
Short Slurs
No Moderation
A Little Moderation
Some Moderation
More Moderation
A Lot of Moderation
Using Disability as an Insult
No Moderation
A Little Moderation
Some Moderation
More Moderation
A Lot of Moderation
B. Level of Moderation
Design 2C: Intensity of Ableism
Choose Your Own Moderation Setting:
Short Slurs
Nothing
Mildly Ableist
Ableist
Somewhat Ableist
Very Ableist
Using Disability as an Insult
Nothing
Mildly Ableist
Ableist
Somewhat Ableist
Very Ableist
C. Intensity of Ableism

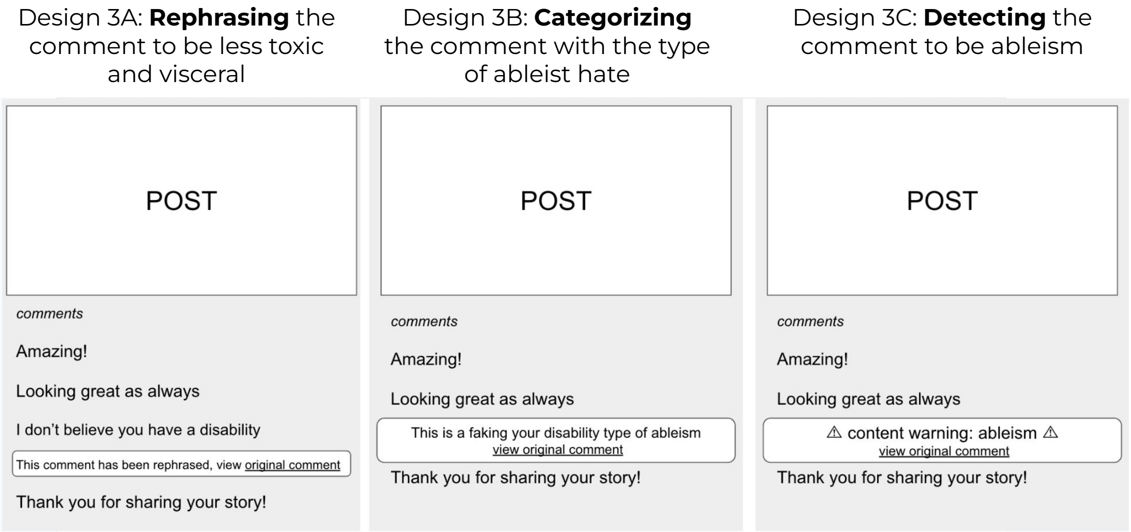
Figure 3: Diagram of Design Probe 2, configuring filters based on ableist types of hate. This includes 3 different designs: 2A (toggling each ableist type), 2B (slider for quantity of each ableist type), and 2C (slider for intensity of each ableist type). To reduce cognitive load, we asked participants to imagine design 2B and 2C to be applied to all the ableist types.

view such filters. We sought to explore how disabled people perceived filters specifically aimed at addressing ableism.

We used design probes to explore future designs spaces of using personalized moderation to address ableism. We shared probes related to: 1) designs to configure ableism-specific filter settings (Design Probe 1 & 2) and 2) designs on how the tool acts on these settings; for example instead of filters fully removing the content we explore other alternatives like rephrasing the hateful comment or a content warning (Design Probe 3).

Filter Configuration (Design Probe 1 & 2). The first set of design probes featured varying filter interfaces for configuring user preferences. Design 1 (see Figure 2) allows users to configure based on ableism, similar to current toxicity scales [47]. Design 2 (see Figure 3) allows users to configure based on types of ableist hate, using Heung et al.'s taxonomy of ableist hate and harassment [42]. Within Design 1 and Design 2, we presented varying control elements, similar to Jhaver et al.'s [47] personalized moderation designs for toxicity. This included: A) toggle (on/off functionality), B) a slider on the proportion of moderation (percentage of ableist posts randomly removed), and C) a slider on the intensity of ableism. These design probes provoked preferences on ways to configure given levels of granularity and labor as well as participants' overall perceptions of using AI to filter ableist content.

Presentation of Hate (Design Probe 3). The third set of design probes (see Figure 4) explored personalizing the presentation of the hate, specifically augmenting the visibility of the hateful comment. Prior work has found social media users are hesitant to set restrictive filters due to their fear of missing out [47]. Furthermore, disabled users, particularly creators and advocates, may want to


Design 3A: **Rephrasing** the comment to be less toxic and visceral
Design 3B: **Categorizing** the comment with the type of ableist hate
Design 3C: **Detecting** the comment to be ableism
POST
POST
POST
comments
comments
comments

Amazing!
Looking great as always
I don't believe you have a disability
This comment has been rephrased, view original comment
Thank you for sharing your story!

Amazing!
Looking great as always
This is a faking your disability type of ableism
view original comment
Thank you for sharing your story!

Amazing!
Looking great as always
⚠ content warning: ableism ⚠
view original comment
Thank you for sharing your story!

Figure 4: Diagram of Design Probe 3, customizing the presentation of the filtered hate. The original hateful comment contained slurs and an accusation that the user was faking their disability. We presented three designs to obscure or describe the hate, in contrast to current filters which completely remove moderated comments: 3A (AI rephrasing of the hate), 3B (content warning categorizing the type of ableist hate), and 3C (general content warning of "ableism"). For each design, users have an option to click and view the original comment.

read hate in order to make new advocacy content. We thus presented three design probes with varying levels of visibility to the original hateful comment. This included: A) rephrasing hate *(highest level of visibility of the original comment)*, B) a content warning categorizing the type of ableist speech, and C) a content warning detecting ableism *(lowest level of visibility of the original comment)*. The content warning design was inspired by prior work which found that trans users preferred customizable content warnings on social media [36].

3.3 Analysis

Two of the authors analyzed the data by first preparing the transcriptions, listening to the audio recording to fix any issues with the automated transcriptions. We coded one interview and one focus group separately, then came together to discuss the codes. We repeated this process again with another interview and focus group recording. Through collaboratively discussing codes, we settled on a preliminary codebook, which included descriptive codes and codes tagging which design probe participants were responding to. We then split up the remaining transcripts, discussing new codes, consolidating codes with similar meanings, and collaboratively refining the codebook into categories (e.g., "concerns" and "wants"). We then conducted thematic analysis [16], clustering similar codes together on a digital board and finding themes across categories of codes (e.g., "ableism is ambiguous" and "viewing hate is important for safety"). All authors participated in peer debriefing as interviews and focus groups were conducted and gave feedback on the codebook, clustering of codes, and themes. The iterative discussions throughout data collection and coding led to consensus.

3.4 Positionality

When conducting research on online harms, especially given the disproportional effect of hate and harassment on historically marginalized populations, it is essential to reflect as researchers. All authors have experience conducting research with and for the disability community and some members of the research team are disabled. We value amplifying the perceptions and experiences of disabled people within conversations on platform moderation and online safety. While personal moderation tools can be beneficial in mitigating harm caused by viewing ableist hate, such tools are not a solution to ableism, nor do they absolve a platform of their responsibility to remove harmful ableist content.

4 Findings

In this section, we share themes from the interviews (RQ1) and focus groups (RQ2). First, we present participants' experiences addressing ableist hate and harassment online. We highlight their experiences using existing personal moderation tools, including blocking and word filters (Section 4.1). Then, we share participants' perceptions and preferences on the design of an ableism-specific

AI filter (Section 4.2 and 4.3). Throughout these sections, we interweave their perceptions on AI filters and their capability to identify ableist text. Lastly, we share how personal moderation has limitations in effectively reducing the harm of ableist hate (Section 4.4). Throughout our findings, we note which participants identified as content creators (C#).

4.1 Addressing Ableist Hate & Harassment Online

4.1.1 Responding & Educating. Some participants responded to perpetrators of ableist hate by educating, especially if participants' thought it was rooted in genuine ignorance, *"a misunderstanding,"* (C18) or a result of someone being *"uninformed"* (C9). However, participants were also wary of the risks of educating, which could escalate the hate further. C12 explained how publicly educating someone led to further harassment from others:

> *"I was trying to be helpful by informing that it would be best to avoid that term [hearing impaired], because most people in the deaf community prefer... deaf or hard of hearing. And that person said, 'okay thank you.' But other people came in saying,' that's a lie', 'that's not true'... [and] DMs saying, 'oh, you're hearing impaired... Did I hurt your feelings? Lol.'"*

A few participants responded to ableist hate with humor and satire. C18 described how she used her *"vulnerability and sense of humor"* to effectively educate. C18 recalled an instance when someone said to her *"why don't you just stay at home?"*, and in response C18 created a post showing her accomplishments and *"out having fun"* as a way to *"reclaim it."* C23 explained that she uses humor to showcase her resilience towards hate.

> *"I try to embarrass them and it's kind of funny... He's like, "I'm sorry no one wants your crippled p*ssy" and my response was, "that's not what your daddy said, so you need to stop before I give you another sibling and cut you out of the will.'... I just need people to know that I don't take this seriously."*

On the other hand, many participants explained that responding directly to harassers was not worth their energy and likely will not be productive. A few participants felt that they do not have the *"bandwidth"* (C22) to be involved in sustained dialogue with harassers, which frequently led to *"comment wars"* (C9) with no resolution. Furthermore, some participants refused to respond to harassers due to the social media algorithm giving hate more engagement, especially when *"some accounts thrive on interacting with ableism"* (C7). C11 explained that responding could even benefit harassers because *"it's giving them engagement"* and *"the fact that they've got a response means that they're probably more likely to... do it to other people who aren't as good as dealing with it."*

4.1.2 Blocking. All participants have blocked harassers before, noting it to be *"the best option for your own health"* (C9). Blocking provided a form of relief to participants. P21 explained that by blocking *"you're addressing the issue... [the harasser] can't see me in any content or situation, and vice versa, and that's great."* Participants also shared that blocking stopped harassers from *"sending harassing content"* (C14) and helped participants *"avoid future hateful comments"* (C5).

Participants also shared ways blocking could prevent hate from other users beyond the harasser. For content creators, blocking harassers prevents them from *"negatively influencing [the creators'] followers,"* (C5) and prevents hate from escalating. Another participant explained that blocking could help *"control the reach of posts"* (P13). P13 would block someone if a harasser quote tweeted about him in order to stop the post from reaching their harassers' followers, preventing potential hate from other users.

On the other hand, some participants explained the downsides of not viewing the blocked person's content, especially if the harasser was *"talking about [them]"* and saying *"things that are not true"* (C16). C5 explained that blocking meant they could no longer educate them on disability. C11 admittedly unblocked a harasser to inspire new advocacy content: *"I have unblocked someone... to see what content I could create to almost combat it, which maybe isn't the healthiest thing... [but] I know that other people are seeing this, and I want there to be another side to the argument."* Creators may choose to unblock or refrain from blocking harassers in order to counter ableist hate and engage in advocacy, despite the potential emotional or psychological toll it may take.

Participants also shared *"loopholes"* (C14) harassers used to circumvent being blocked. Participants mentioned harassers could avoid account bans by creating new accounts and could *"avoid the IP ban"* (C12) by using a VPN to modify their IP address. C14 expanded on harassers' strategies of creating new accounts: *"there's an account called 'I hate wheelchairs'... They keep creating new accounts [and] they're on 'Ihatewheelchairs4'... I don't understand how Instagram... allows that username to be made."*

While most participants felt blocking resolved hateful interactions, a few participants shared that blocking their harasser may escalate the hate from other social media users. This was especially worrisome for creators who had harassers with a large following. For example, C17 expressed a concern of blocked accounts screenshotting that they were blocked as a *"flex,"* making it likely for *"their followers to harass you"*.

Other participants expressed additional features that would enhance their experience with blocking. C9 and P15 requested a *"block and remove interactions feature"* where they could block and remove all public interactions with that account, including posts the blocked account had liked or comments on posts. They suggested this feature would avoid being triggered by the blocked account's username, when looking through past posts. P21 desired a feature that would allow them to block someone without completely removing them from the Facebook group. They explained that the blocked individual may still have the right to be part of the group, but P21 wanted to block that account for her own "safety and sanity."

4.1.3 Word Filters. The majority of participants did not have experience using word filters, and of those who had used word filters most of them were creators. Some participants who used word filters inputted phrases to remove ableist hate, such as the r-word (C17, C12) or *"faker"* (C14) to account for accusations that they are faking their disability. C14 also input her personal information into these systems to avoid being doxxed[4]. She detailed how using word filters was laborious: *"I'm constantly updating [my word filters] and adding different variations ofa... annoying ableist phrase."*

A few participants used word filters for non-hate related moderation (C11, C16, P21), like spam. For example, C11 explained that when she made a post with #chronic-illness, she received spamming comments invalidating her disability identity, saying *"doctor so and so can heal you with his herbal blah."* C11 then added "herbal" to her word filters; however, she was wary that genuine comments related to herbal tea could be filtered out.

Participants who used word filters noted flaws in the system, including filters hiding non-hateful comments while failing to block hateful comments. For example, C12 used a preset moderation filter on Twitch, and found it to filter out words he did not feel were offensive. While C12 noted the false positives, C16 entered keywords of a repeated comment she was getting from a pornographic account, but the word filters were not effective in stopping the account from commenting. C16 blocked some of those accounts but she *"can't block them all off [because] it's a lot of work."*

While most participants shared experiences on mainstream social media platforms, C9 described her experiences of Tinder's version of a word filter. She liked the design of first identifying potentially inappropriate words and then prompting the user if they want to block and report.

> *"If a guy... says 'Can I see a nude photo?' Tinder will ask you if the message makes you uncomfortable and if you say yes, it reports the message and blocks the dude... I like that cause it tells me right away, it's okay for you to report this... [and]* it gives you the ultimate say, 'are you uncomfortable?'... I like that it gives people agency."(C9)

4.2 Perceptions on AI Filters Identifying Ableism

In response to Bleep's interface, Design Probe 1 (configuration via ableism), and Design Probe 2 (configuration via ableist types), participants shared their perceptions and preferences on configuring AI filters to identify ableist text. Furthermore, participants revealed their concerns on AI's capabilities of correctly identifying ableism, shared implications of AI's inaccuracies, and brainstormed design suggestions to account for AI's inaccuracies.

4.2.1 Preferred Configuration Design: Ableist Types. Overall the majority of participants preferred configuring their filters via ableist types (Design Probe 2) because it was granular and understandable as to what the filter would be removing. Before viewing Design Probe 2, a few participants suggested a similar design that allowed them to filter out specific forms of ableist speech. P13 suggested a checkbox design that could allow him to input his preferences based on what he defines as most hurtful.

> *"Rather than a sliding scale, boxes that you could check for... would be more helpful. Because patronizing comments... get to me more than just straight up hate speech cause I can just... block... But the patronizing comments... [I have to] engage with you now, teach you something... they can take a lot more time and energy...They [the comment] might be not as ableist, but they might not be the lowest impact."* (P13)

In response to Design Probe 2, participants shared how they would configure their filters, highlighting how some types of ableist speech are more emotionally draining than others. For instance, C9 prefers configuring by ableist types, because she can filter with a level of specificity that aligns with her personal triggers and preferences.

> *"What matters... [is] the actual content of what's being said... If someone makes a patronizing comment to me, I'm just gonna make a joke back... but... I really really feel uncomfortable when I'm doing a nail tutorial and I have people saying sexually fetishizing things, because, A) my parents follow me... B) I just don't want to see it, because it makes me not want to post photos of myself... I sit there and think is that what everyone is thinking of me?... It's the intent of what is said that matters... it makes me feel unsafe."* (C9)

For this very reason, the majority of participants expressed ableism (Design Probe 1) to be too broad of a concept, too vague, and expressed hesitancy to use the filter because they were unsure what was being filtered out. For example, after viewing 1C (intensity slider of ableism), C18 explained *"I don't think that... [an intensity]* scale... would necessarily accomplish...instances [when] I do wanna see less of something." Conversely, a few participants were comfortable with the vagueness of "ableism" and preferred a straightforward method for filtering out ableist speech. For example, P2 appreciated Design Probe 1A, which allowed him to simply *"toggle it on and [he's] good to go."*

4.2.2 Preferred Control Elements. Participants shared their varied preferences for control elements (A: toggle, B: slider of moderation percentage, C: slider of intensity) for each configuration design (Design 1: configuration based on ableism, Design 2: configuration based on ableist types).

Toggle vs. Sliders. Overall, participants preferred toggles over the sliders. Some participants liked design 1A, since it was *"easy to navigate"* (P4) and they did not want to view any sort of ableism. The majority of participants wanted more control over what types of ableist content they wished to not view, so they felt Design 1A was*"too vague"* (C11). Therefore, they preferred using toggles to configure via ableist types (Design 2A) as that allowed them enough granular control to decide whether or not they wanted to view that type of ableist speech. C23 expanded on why she preferred an on/off mechanism:

> *"I won't have to battle within myself over the intensity of what I want to see... it's just all or nothing... [Do] I need to face reality?... What days can I accept this harassment and what days do I want to fight it?"* (C23)

Similarly, C10 elaborated on why a toggle may be more fitting for certain types of ableist hate.

> *"I hate when people say I'm faking a disability or... that I don't 'look disabled.' I would not want: slightly, mildly, a little, [or] a lot. I would want none at all... I feel like the sliders don't make a lot of sense, because... why would I*

[4] being harassed by revealing someone's private information without their consent

> *want to let some of this still come through when it's the most upsetting thing to me?"*

Several participants also noted that a toggle was more accessible for those who use *"screen readers or who might have hand dexterity issues"* (C11).

Moderation Slider (Design 1B/2B) vs. Intensity Slider (Design 1C/2C). While some found "percentage of moderation" to be an easily understandable and objective measurement, a majority of participants felt that it did not make sense for mitigating harm. P21 explained: *"it's just randomly selecting 25% or 50%... that doesn't seem like it's... moderating."*

Participants generally preferred the intensity slider over moderation, but several participants also expressed flaws with the intensity slider. In the focus group, C7 pointed out that one *"cannot quantize disability, accessibility, or ableism."* Using humor, the focus group attendees explained how quantifying ableism minimized and trivialized the harm done. C7 said: *"mildly ableist or somewhat ableist sounds weird to me. It sounds like a joke...'hey, someone punched me in the face.' And they were like, 'yeah but did they mildly punch you, or did they really punch you?"*

Other participants also explained that *"degree of ableism"* (C18), referring to Design Probe 1C, does not allow them to pick the content they personally find hateful. This is why many participants preferred ableist types (Design Probe 2).

4.2.3 What is Ableist is Contextual, Equivocal, and Contested. Participants were skeptical if AI filters could accurately identify ableism due to 1) the subjective nature of what is ableist, 2) the nuance and context of when something is ableist, and 3) perceived biases and capabilities of AI.

Participants questioned if there is a universal understanding of what is ableism or what is ableist. This was a common concern, especially when configuring filters by ableism (Design Probe 1). Participants were unsure what types of content would be filtered out. C9 explained that ableism is *"a broad category... not everything is the same level of problematic to each individual person."* Participants shared examples of how, even within the disability community, there are disagreements on what is ableist. Participants referred to diverse preferences regarding self-identifying language (e.g., people with special needs, disabled people vs. people with disabilities, hearing impaired vs. deaf and hard of hearing). C7 and C16 explained that if AI were to block words related to one's identity, this could cause fear of being *"penalized"* for your own identity:

> *"I have quite a few friends who identify as hearing impaired and impaired is a very disgusting word in the disability space. But if she identifies that way, they are allowed to... I worry that those types of voices will be moderated out."* (C16)

Since ableist language is not clear-cut, C11 explained that: *"language that might be considered offensive.... also might exist for a good reason... I think that's just important to note when you're training AI, [ableist language is] just not explicit or definitive."*

Participants also viewed ableism as intertwined, but not identical to body shaming. Intel's Bleep interface combines ableism and body shaming into one category of hate. C9 expanded on why this design is problematic.

> *"As someone who is both fat and disabled, [ableism and body shaming] are different issues... not all disabilities are physical or has to do with the body... It feels like a very narrow definition of disability when you also loop it with body shaming, because to me it looks like they're only looking at visible disabilities."*

4.2.4 Skepticism and Concerns with AI Accuracy. Since ableism is subjective and nuanced, many participants were skeptical if AI could accurately identify ableist comments. Participants were concerned that the filters would take keywords out of context, leading to false positives (AI wrongfully filtering out comments about disability). C19 explained that she does not *"trust AI... to distinguish what is ableist and what is disabled or disability adjacent"*(C19). This included filtering out ableist terms used within historical, medical, educational context that are not intended to be hateful towards an individual, or if a user was recounting a hateful experience and directly quoted hate they had received. The potential of false positives caused participants to fear missing out on constructive conversations and opportunities to connect with the disability community. For example, C9 expressed uncertainty of how well AI can understand context:

> *"Would you not be able to read historical things because they use words differently? If you're talking about a... time when... the [r-word] was acceptable... I feel like without there being a contextual component you're gonna miss, so many conversations that may not be what this thing [AI] thinks it is, but... I don't understand AI that well."*

Similarly, participants were also concerned with AI filtering comments with disability-related reclaimed words (e.g., cripple, gimp). Participants agreed that what differentiates a slur from a reclaimed word was dependent on who said it and the intent. P13 questioned how AI filters could account for this context:

> *"There's the tension within in-group and out-group language... not everybody has earned the right to use certain words. But within a group... there's safety and familiarity... One word that I would be really upset to hear from somebody who wasn't also disabled would be cripple. But that's a word I've definitely used... And it's a word that ... has scholarly usage... So how is the intent... accounted for?"*

A couple of participants anticipated the AI to be inaccurate due to ableist biases, assuming that AI models may not be trained on the lived experiences of disabled people. P13 anticipated the AI to over-filter content about disability and sex, because of stereotypes of disabled people being desexualized. Furthermore, some participants doubted AI's accuracy due to negative experiences with online moderation. Creators were especially wary given their past, negative experiences with moderation.

> *"When I try to post content... it'll be reported as hate speech because I'm addressing someone as an 'able-bodied Savior or a white Savior'... These AI filters end up blocking a lot of disabled creators from sharing very important information... some of us feel like we've already been burned by moderation."* (C23)

Over time, this may *"inadvertently chill or silence the speech of pro-affirming disability language like conversations about disability, or about diversity, or about bodies in general" (C18).*

4.2.5 Designing for Inaccuracies. Implications of AI inaccurately removing disability-related content included missing out on constructive conversations and infringing on disability activism and mobilization. Participants viewed filters as a possible threat toward the community they built online. For example, C23 explained:

> *"I'm afraid that I'll miss things, and it'll harm relationships that I've built...disabled people have found kinship on social media. Because for a lot of us we're the only disabled person we've ever known. We live in communities that can be isolating. So I don't want to ever miss [out] because of a filter that I've set."*

Participants proposed additional features to alleviate these concerns. Several participants wanted a feature to select certain accounts to be exempt from the filter setting. This way, participants would not miss out on their disabled friends' posts, knowing that if they did use an ableist term it would not be triggering or considered hateful. C7 explained: *"let's say C11 and I are interacting on Twitter, and I know that she uses the word cripple to self-identify. I would put her on 'no moderation' or 'a little moderation' if I know... the word cripple is going to be triggered by [the filter]."*

Since participants were unsure how AI is defining ableism, they wanted oversight to how it was filtering content. This included viewing what the filter is removing to *"determine, is this worth it? Is it working?"* through a *"test button"* (P21). Some wanted an *"undo button"* (P3, C22) to override any mistakes the AI makes, such as filtering out a friend's comment that they would have otherwise responded to.

Some participants wanted to train the AI to align with their preferences by tagging content they found as ableist.Participants also explained that it was important to communicate to the algorithm not only what is ableist vs. what is not, but also why.

> *"If AI learns from many actions, then the more information we can give it, probably the better. So if one is accepted or allowed through [the filters], the why or the justification could be: 'this is a person in this community with this identity has chosen to reclaim this language.' And one that gets rejected might be 'this is a person… harassing someone'... I would probably type 2 or 3 sentences... [but] I don't know if the AI would understand that."* (C20)

C18 added that teaching the AI allows each user to *"define for themselves what ableism means and looks like."*

4.3 Varied Tolerances on Viewing Ableist Hate

In response to Design Probe 3, participants shared their own individual preferences of if (Section 4.3.1) and how to view ableist hate (Section 4.3.2 & Section 4.3.3), acknowledging that other disabled users have *"different levels of tolerance"* of viewing ableism (P13).

4.3.1 To View or Not to View Hate. Participants shared their philosophies of whether or not to view hate. A couple of participants prioritized protecting themselves from viewing hate online, wanting the ability to turn off all hate. For example, P3 explained that *"we are responsible for our own safety... I would rather just toggle the entire thing out."*

Some participants explained that, while in theory they wouldn't want to view hate, in practice they would be too curious to not. Although participants appreciated having the option to view the original comment in Design Probe 3, a majority of participants acknowledged that having an option to *"view original comment"* was too tempting to not click. C17 explained:

> *"I'm just always so curious... I would probably say 80% of the time. It's just gonna bug the hell out of me if I don't know what is being said, because… I need to have that control. But then there's gonna be that other 20% [where] I just don't feel like it today."*

Due to inherent curiosity, a couple of participants wanted the *"view original comment"* button to be removed.

Several creators stressed the importance of knowing what is being said about them in order to *"control the narrative"* and felt a strong responsibility to moderate their own page (C9). C9 explained that if other users are spreading rumors and talking about him on his own channel, then *"ignorance is not bliss."*

While hate is harmful, a few participants emphasized the importance of viewing hate for their own physical safety. C23 was concerned about hiding death threats: *"if it's a serious threat, someone should be notified, whether it's [the] police or anyone."* P13 highlighted the trade-off of protecting themselves from hate and being aware of potential dangers.

> *"As much as I don't like seeing hate speech, it is helpful to know the conversations that are happening right, what kind of is the Zeitgeist... and to just automatically screen that out also doesn't really contribute always to a sense of safety. Because it means you actually can't be aware of potential dangers"* (P13)

These participants highlighted the need for personal moderation tools to balance protection from hate with awareness of serious risks, such as death threats.

4.3.2 Rephrasing Feels Fake, Patronizing, and not Helpful. The majority of participants' initial reactions towards rephrasing was negative, finding AI rephrasing hate as uncomfortable and dishonest. C9 explained how the rephrased version of hate *"feels fake… I'd rather know what somebody just said so then I can report it."* (C9) Furthermore, several participants discussed how rephrasing hate felt *"patronizing [and] infantilizing"* as if participants needed protection from AI to give them the *"nice version"* of hate (C7). A few participants were skeptical of AI's ability to generate an *"authentic translation"* and rephrasing of the original comment (C19).

Participants also added that it's not worth rephrasing hate if everyone else online can still view the original hate. P8 explained that *"it [3A] doesn't really make sense to see a softer version of [an] insult and everyone knows in the comment section it's saying something else... I don't feel it's necessary. If it's hateful, let me see it that way."*

Many participants explained that rephrasing hate was largely not effective, since it does not change the intent of the harasser. C23 called rephrasing ableist hate as *"diet ableism, it's diluted [and] it'll hurt less"* but *"it's not sparing anybody's feelings."*

Several participants suggested the rephrasing function should instead be used to educate harassers. C18 explained that while rephrasing helps with *"softening people's hatred,"* she feels it reduces accountability by *"letting [harassers] off the hook for their hateful language."* Additionally, P21 clarified why she rather have the system to be used to educate harassers:

> *"[3A] seems like a strange way of trying to protect our feelings, which being disabled, I don't want you to try and protect my feelings like I will take actions to do that... if everybody else can still see the damaging words, then… What is the point of this? "*

4.3.3 Content Warnings Are Informative. The majority of participants preferred content warnings (Design Probe 3B & 3C) before viewing ableist hate, though their preferences on the amount of detail about the hate varied. Some participants preferred 3B (categorizing hate) because it's *"helpful… to get to know the information [about the hate] but not feel attacked"* (C5).

Categorizing the hate empowered the user to make an informed decision of whether or not to view it. Since the comment is not removed, the user has agency to view it at any time. For example, C16 said *"I like that it gives me the option to look at it or not, and I can decide based on my mood at the moment… so I can look at it later, when I can handle it."* C7 emphasized the importance of designing for autonomythat was missing when AI rephrased hate for them.

> *"Many disabled people have that autonomy taken [away] from them, or are told they don't know how to make decisions or patronized... [a content warning]is more like a trigger warning where you are given the autonomy, whether you want to read it... That's not taken away from you as it was in [Design Probe 3A]."*

A few participants added how categorizing the hate improved the explainability of the filter system. C10 described how the category is not only *"giving an explanation as to what kind of comment was said"* but also explaining *"why the AI detected and flagged the comment."* C11 elaborated that the explainability built more trust with the AI and reduced the likelihood of viewing the original comment.

> *"If it's just "ableism" (referring to Design Probe 3C) again for that morbid curiosity thing, you might want to read it, or you might want to check whether it actually is ableist or not, whereas if it's more specific... there's a little bit more trust that the AI knows what it's doing if it can detect the category, whereas if it's just ableism, the trust is less."*

On the other hand, some participants preferred a general content warning (Design Probe 3C), because it is enough information to make a choice of whether to view the content. A few participants added that categorizing ableist hate (Design Probe 3B) may be too triggering. For example, C16 said: *"I have mixed feelings about announcing the type of ableism, because… on a bad day I don't even wanna know that it was 'your faking disability' type of ableism."* With a similar sentiment, C5 and C12 preferred Design Probe 3C since it was a generic warning with less risk of being emotionally triggered.

4.4 Limitations of Personal Moderation

While some participants shared how filters have the potential to reduce harm of viewing ableist content, other participants wanted filter settings to be applied to everyone's view, not just their own view. Having the hate viewable for others may cause harm to others or escalate hate. Participants felt that inadequate moderation for ableism seemed unfair since other types of discrimination were being removed by the platform.

> *"Does it [the filters] protect me more? Yes. But if people can engage with it, comment on it, or view it potentially. I don't like that... Other hate speech that is more widely recognized is automatically deleted most times on other posts so why is Ableism an exception."* (C14)

Participants added how the various presentations of the hate (Design Probe 3) could be educational and therefore wanted all social media users to view it. More specifically, participants shared that the filters based on ableist types and the detailed content warnings could spread awareness of the different types of ableism. For example, P21 liked the ableism content warning and wanted everyone to be able to view the content warning as a way to educate others that ableism exists. P21 explained how she imagined a teaching tool to be designed using all of the different versions of Design Probe 3.

> *"[Design Probe 3] should lead with... 'content warning: ableism', and then underneath it say[s] 'this is the faking your disability type of ableism.' And then under that 'this comment has been rephrased to say, I don't believe you have a disability.' And then underneath that it would say, 'view original comment.' So then you're teaching everyone...[other users are] learning... what is ableism,... the type of ableism, and... how to rephrase something."*

C18 added that the original poster should be notified if their content is filtered out by another user and *"tagged as this type of ableism,"* so the harasser can be educated on why it's ableist.

Other participants acknowledged the limitations of filters: it does not hold users accountable for perpetuating ableist speech. Therefore, a few participants recommended the filter system to trigger repercussions like a suspension. P13 explained that *"personalized moderation isn't really a solution to unsafe communities and unsafe spaces… [There] needs to be a community responsibility as well."* Participants felt that *"community responsibility"* needed to come from the platform moderation itself by seriously addressing reports of ableist speech. C23 attributed lack of moderation of ableist hate due to lack of knowledge on ableism:

> *"Harassing disabled people is the one thing you can get away with pretty easily on social media... Attacking someone's disability isn't seen as a problem because most of the moderators aren't disabled. They're never going to challenge comments like, 'you shouldn't be proud to be disabled'... [because moderators] themselves have those same beliefs."*

5 Discussion

While personal moderation enables users to control for various types of content, ranging from content that is harmful to content that is uninteresting, our study examines one specific type of content users may prefer to avoid: ableist hate. Given this context, we present design recommendations for an ableism-specific AI filter to support safety, harm reduction, and agency. We also make recommendations related to usability, explainability, and trustworthiness of the system, as it may impact whether or not users adopt AI filters. Lastly, we share our study's limitations and directions for future work.

5.1 Design Recommendations for Personalized Moderation

Our design probes elicited participants' values when using a personal content moderation tool during their experiences with ableist hate online. We discuss design recommendations that support values participants' shared (refer to Table 2 for a summary).

5.1.1 Threat Notifications Promote Safety. Since personal moderation tools are designed to protect users online, it's essential to address situations where they may unintentionally create other safety risks. As noted by our participants, while personal moderation might enhance psychological comfort by removing ableist speech from view, it can compromise safety by reducing a user's awareness of potential dangers.For example, moderation tools may completely filter out hate related to physical safety, such as death threats. Tune, Google's content moderation tool, has a disclaimer of this limitation: "Tune isn't meant to be a solution for direct targets of harassment (for whom seeing direct threats can be vital for their safety)" [9]. Removal of threats to one's safety may lead to ignorance of unsafe physical spaces and events and lead to in-person harms [82]. This is increasingly relevant as social media has become a central information hub for news and events [4]. Additionally, eliminating ableist hate might obscure the cultural climate or "zeitgeist." This may lead to disabled individuals being unaware of potential risks of harassment related to disclosing their disability [26, 42, 43] and emotional harms to their self-esteem and possible internalization of disability stereotypes [75].

Removing ableist speech may also reduce the agency of disabled people. This study and prior work has shown how both disabled and non-disabled social media users alike use blocking, reporting, and responding strategies to address hate [26, 39, 42, 43, 67, 78, 89]. Our participants were hesitant to use filters, given that it could prevent them from using blocking, reporting, and responding to reduce anxiety and prevent hate from escalating. For example, blocking prevented repeated harassers posting ableist comments. Content creators or high-profile users wished to control narratives about themselves, as leaving hate could cause reputational damage [89], lead to more engagement and hate from other users due to the platform algorithm [13], and subsequently may escalate into in-person harms [61].

Not accounting for the above use cases may lead to increased safety risks towards users from historically marginalized communities who are at-risk of identity-based hate. We recommend for personal content moderation systems to consider not removing threats to one's physical safety, but instead design notifications that inform users of safety concerns without burdening users who are targets to hate [88]. Researchers should explore AI's role in effectively implementing safeguards for users' safety when using personal content moderation tools.

5.1.2 Ableist Types Enhance Explainability & Usability. Current personal content moderation tools often use broad categories like "toxicity" or "sensitivity," which may lack clarity as to what exactly the system is filtering. Prior research [47] highlights the need for more explicit definitions of what constitutes toxic and sensitive context. However, our findings suggest that definitions alone may not effectively communicate what is being filtered. The definition of "what is ableist" is hotly debated within the disability community, complicating how a system *should* define ableism explicitly and how a system should categorize the varying degrees of ableism (e.g., mildly ableist vs. very ableist). Intensity of ableism (design C) may trivialize ableist speech and may invalidate individual perceptions of how ableist a comment is to them. Ableist types can be an alternative to communicate explicit forms of ableist speech without offending or invalidating the user, especially since our findings imply that how people experience ableism is subjective.

Detecting toxic language, either as a binary value (i.e., is it toxic or not?) or by measuring the intensity of toxicity, are common measurements and standard practice in machine learning-based moderation systems (e.g., Perspective API [2]). However, prior work has highlighted biases in toxicity detection systems [31, 34, 70, 80], including ableist biases [40, 91]. Given these existing concerns, we argue that personal moderation requires a shift from classifying hateful text based on toxicity levels to classifying it based on the type of hate found within the text (in our particular context, the type of ableist hate). Participants understood ableist speech as types, not as a numerical value. Enabling disabled people to choose ableist types may be more explainable and usable in configuring what kinds of ableist speech they wish to filter out. For example, some participants found patronizing comments more exhausting than outright ableist slurs and preferred not to see patronizing comments. If a system only affords the user to select severity of ableism, it may not account for personal preferences the user wishes to view and not view. Furthermore, designing the interface with these types may be more intuitive than using a slider that adjusts the intensity of ableism. Allowing users to configure filters based on ableist types aligns better with their preferences for filtering specific content they personally find to be more ableist or hurtful.

5.1.3 Content Warnings Support Agency & Reduce Harm. Similar to prior work on the usage of content warnings on social media [36], our findings suggest users may benefit from content warnings that support informed decision making and control. For example, users could decide to view ableist hate if they were in the "mood," did not find the type of hate particularly triggering, and/or if they wanted to respond to the hate. Some appreciated the explainability of content warnings for categorizing the type of hate as this fostered greater trust with the AI system accurately identifying ableist hate. Participants noted that this clarification improved transparency, a highly valued characteristic among social media users regarding moderation tools [47, 60]. Additional transparency also alleviated concerns of missing out, giving more assurance that the AI was filtering ableist hate, not disability-related content.

Although content warnings were perceived as beneficial, it is important to consider the potential side effects of content warnings.

Table 2: Summary of Design Recommendations for Personal Content Moderation.

Value [5]	**Design Recommendations**
Promoting Safety	Design notifications pertaining to threats to one's physical safety of the user rather than removing them from view
Enhancing Explainability & Usability	Use ableist types for configuring personal content moderation settings
Supporting Agency & Reducing Harm	Embed content warnings as an option to support users' decision making in deciding whether or not to view hate
Building Trust	Implement ways to oversee filtering, undo-decisions, and support AI learning from the user

While content warnings "allow those who are sensitive to these subjects to prepare themselves for reading about them, and better manage their reactions" [62], prior work has detailed how content warnings can backfire. Content warnings may not be helpful for those experiencing trauma or re-traumatization [85]. Content warnings may cause a "forbidden fruit effect," [17] making the content seem more attractive. Participants anticipated this effect by explaining they would be "too curious" to not "view original comment" in Design Probe 3. Future work is needed to evaluate the varying designs and effects of content warnings on social media, especially for identity-based harmful content.

5.1.4 Oversight & Reversibility Features Build Trust. Due to prior negative experiences with moderation, participants expressed skepticism and hesitation towards using AI filters. While previous research indicates that users are wary of AI filters over-moderating and fear of missing out on content [47], our participants raised additional concerns specific to the lack of widespread understanding of ableism. Furthermore, ableism is often poorly moderated; prior work has noted instances of wrongful removal of disability-related content [42, 59] and instances of ableist hate not being addressed by platform moderation [42]. This aligns with research showing that existing mistrust in a domain extends to AI-based solutions (i.e., mistrust in moderation extends to mistrust in AI-based moderation) [56]. This skepticism may be applicable to other historically marginalized groups who have similarly felt unsupported by platform moderation, such as black people [38, 39, 67] and LGBTQ people [38, 66, 70]. Consequently, those with negative experiences and distrust in current moderation systems may also view AI moderation tools with similar suspicion. This is particularly relevant for active social media users, such as creators, who are more likely to encounter negative moderation experiences.

Participants perceived filtering of identity-based harmful content by AI to be risky, especially due to concerns about AI over-moderating content highly relevant to their identity and daily life. Prior work suggests that mistakes by AI filters may be detrimental to disability advocacy, community building, and information gathering [15, 42, 54, 73]. Because of these concerns, our findings suggest that users may be reluctant to adopt AI filters without robust safeguards and fail-safes. Therefore, we recommend implementing features like the ability to undo and correct filtering errors so the underlying model can learn from these corrections. Additionally, to give greater oversight and control over *who* the AI filters out, we also recommend implementing allowlists [24], a feature that allows users to select trusted accounts exempt from filtering. By adding such features, users may feel more confident in using AI filters.

5.2 Limitations & Future Work

Since we focused on showing design probes that were meant to provoke participants' wants [92], we did not build an interactive prototype. This trade-off was intentional, as an interactive prototype may bias participants to only share what they think is feasible [37, 77]. The technical feasibility of an ableism-specific AI filter may not be far off. For example, generative AI models like ChatGPT show promise of moderating user-generated content, as researchers begin to evaluate AI's effectiveness in identifying hateful content [58]. Future research should build and evaluate such a tool, which may provide additional insights on user's behaviors over time, usability considerations, and AI's accuracy in identifying ableist speech. Furthermore, future work should investigate how the usage of personalized moderation varies across different user types (e.g., content creators vs. casual users) and among users with varying identities (e.g., disability, race, sexuality). As these tools aim to encourage safe participation online, it is important that they are accessible and do not contribute to the labor disabled users exert to be on social media [15, 42, 59, 65, 76, 87]. Furthermore, future work should investigate AI-based personal moderation for other forms of identity-hate and other manifestations of hate like image- and video-based hate.

While AI-based personal moderation can contribute to a fairer and less harmful online environment for disabled users, it is not a substitute for structural changes needed to achieve justice for those affected by ableism [12]. Future research should consider the role of AI-based platform moderation to combat structural ableism, such as designing for disability awareness and education. For example, participants wanted rephrasing to be a platform intervention, nudging perpetrators to reconsider posting ableist hate and learn more about ableism. However, it is critical to ensure that proactive nudges do not enable harassers in alternate ways; for example, actors can be even more toxic if subverting moderation systems becomes gamified [94]. Future work should consider incentivizing positive and prosocial user behavior to prevent ableist hate in the first place, while protecting against the potential adverse effects of proactive nudges.

6 Conclusion

This paper investigates how AI-based personalized moderation can safeguard disabled users from viewing ableist hate on social media. We created design probes to elicit users' preferences for an ableism-specific filter, including ways to filter ableist text (e.g., based on types of ableist hate) and ways to customize the presentation of

[5] Values listed are equally important and represent a design space rather than an ordered list.

hate (e.g., AI rephrasing hate or content warnings). We share design recommendations related to supporting users' safety (e.g., being notified of personal threats), improving usability (e.g., filtering based on ableist types), reducing the harm (e.g., content warnings) and building trust (e.g., undoing filter decisions). Lastly, we further conversations on personal moderation to address identity-based harms, amplifying the perspectives of disabled people using personal moderation tools for ableist hate.

Acknowledgments

We thank our participants for their thoughtful feedback during the focus groups and our anonymous reviewers for helping improve the paper. This work was supported in part by Google, President's Council of Cornell Women (PCCW) and the National Science Foundation.

References

[1] 2021. (2021). https://about.instagram.com/blog/announcements/introducing-sensitive-content-control

[2] 2021. Using machine learning to reduce toxicity online. https://perspectiveapi.com/. Accessed: September 10, 2024.

[3] 2023. r/AutoModerator. https://www.reddit.com/r/AutoModerator/wiki/can_do/. Accessed: September 10th, 2024.

[4] 2023. Social Media and News Fact Sheet. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/ [Accessed 11-09-2024].

[5] 2024. Bodyguard. https://www.bodyguard.ai/en

[6] 2024. How Do I Report Something I See on Facebook? https://www.facebook.com/help/814083248683500. Accessed: September 10, 2024.

[7] 2024. How to block accounts on X. https://help.x.com/en/using-x/blocking-and-unblocking-accounts. Accessed: September 10, 2024.

[8] 2024. How to Use AutoMod. https://help.twitch.tv/s/article/how-to-use-automod?language=en_US [Accessed 11-09-2024].

[9] 2024. Tune (experimental). https://chromewebstore.google.com/detail/tune-experimental/gdfknffdmmjakmlikbpdngpcpbbfhbnp?pli=1

[10] 2024. What happens when you block someone on Instagram. https://www.facebook.com/help/447613741984126. Accessed: September 10, 2024.

[11] Kathryn Zickuhr Myeshia Price-Feeney Amanda Lenhart, Michele Ybarra. 2024. Online Harassment, Digital Abuse, and Cyberstalking. https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/. Accessed: September 10, 2024.

[12] Cynthia L. Bennett and Os Keyes. 2020. What is the point of fairness? disability, AI and the complexity of justice. *SIGACCESS Access. Comput.* 125, Article 5 (mar 2020), 1 pages. doi:10.1145/3386296.3386301

[13] Thales Bertaglia, Catalina Goanta, and Adriana Iamnitchi. 2024. The Monetisation of Toxicity: Analysing YouTube Content Creators and Controversy-Driven Engagement. In *Proceedings of the 4th International Workshop on Open Challenges in Online Social Networks* (Poznan, Poland) *(OASIS '24)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3677117.3685005

[14] Brianna Blaser and Richard E. Ladner. 2020. Why is Data on Disability so Hard to Collect and Understand?. In *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, Vol. 1. 1–8. doi:10.1109/RESPECT49803.2020.9272466

[15] Katya Borgos-Rodriguez. 2021. Understanding and Amplifying Labor among Content Creators with Disabilities. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 241–244. doi:10.1145/3462204.3481784

[16] V. Braun and V. Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[17] Victoria M. E. Bridgland, Payton J. Jones, and Benjamin W. Bellet. 2024. A Meta-Analysis of the Efficacy of Trigger Warnings, Content Warnings, and Content Notes. *Clinical Psychological Science* 12, 4 (2024), 751–771. doi:10.1177/21677026231186625 arXiv:https://doi.org/10.1177/21677026231186625

[18] F. K. Campbell. 2008. Refusing Able(ness): A Preliminary Conversation about Ableism. *M/C Journal* 11, 3 (2008). doi:10.5204/mcj.46

[19] Fiona Kumari Campbell. 2010. Contours of Ableism: The Production of Disability and Abledness. (01 2010). doi:10.1057/9780230245181

[20] Dasom Choi, Uichin Lee, and Hwajung Hong. 2022. "It's not wrong, but I'm quite disappointed": Toward an Inclusive Algorithmic Experience for Content Creators with Disabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 593, 19 pages. doi:10.1145/3491102.3517574

[21] Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. 2022. Personalized Interventions for Online Moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media* (Barcelona, Spain) *(HT '22)*. Association for Computing Machinery, New York, NY, USA, 248–251. doi:10.1145/3511095.3536369

[22] Maeve Duggan. 2017. Online Harassment 2017. *Pew Research Center* (2017). https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

[23] Dana Dunn. 2021. Understanding ableism and negative reactions to disability. https://www.apa.org/ed/precollege/psychology-teacher-network/introductory-psychology/ableism-negative-reactions-disability

[24] Samantha Dunn. 2023. Blacklist & Whitelist: Terms To Avoid. *Splunk* (2023). https://www.splunk.com/en_us/blog/learn/blacklist-whitelist-inclusivity.html

[25] Jared Duval, Ferran Altarriba Bertran, Siying Chen, Melissa Chu, Divya Subramonian, Austin Wang, Geoffrey Xiang, Sri Kurniawan, and Katherine Isbister. 2021. Chasing Play on TikTok from Populations with Disabilities to Inspire Playful and Inclusive Technology Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 492, 15 pages. doi:10.1145/3411764.3445303

[26] Tessa Eagle and Kathryn E. Ringland. 2023. "You Can't Possibly Have ADHD": Exploring Validation and Tensions around Diagnosis within Unbounded ADHD Social Media Communities. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 29, 17 pages. doi:10.1145/3597638.3608400

[27] Maggie Engler. 2022. Middleware and the Customization of Content Moderation. (2022). https://integrityinstitute.org/blog/middleware-and-the-customization

[28] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (Jun. 2018). doi:10.1609/icwsm.v12i1.15033

[29] Ashish Goel Roberta R. Katz A. Douglas Melamed Marietje Schaake Francis Fukuyama, Barak Richman. [n. d.]. MIDDLEWARE FOR DOMINANT DIGITAL PLATFORMS: A TECHNOLOGICAL SOLUTION TO A THREAT TO DEMOCRACY. *Stanford Cyber Policy Center* ([n. d.]).

[30] Carli Friedman and Aleksa Owen. 2017. Defining Disability: Understandings of and Attitudes Towards Ableism and Disability. *Disability Studies Quarterly* 37 (2017). https://api.semanticscholar.org/CorpusID:151902189

[31] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling Bias in Toxic Speech Detection: A Survey. *ACM Comput. Surv.* 55, 13s, Article 264 (jul 2023), 32 pages. doi:10.1145/3580494

[32] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* 1–288 pages. doi:10.12987/9780300235029

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

[34] Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 363 (nov 2022), 28 pages. doi:10.1145/3555088

[35] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law & Technology* 17 (2015), 42.

[36] 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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 124 (oct 2020), 27 pages. doi:10.1145/3415195

[37] Jack Hakim and Tom Spitzer. 2000. Effective prototyping for usability. In *Proceedings of IEEE Professional Communication Society International Professional Communication Conference and Proceedings of the 18th Annual ACM International Conference on Computer Documentation: Technology & Teamwork* (Cambridge, Massachusetts) *(IPCC/SIGDOC '00)*. IEEE Educational Activities Department, USA, 47–54.

[38] Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T. Hancock, and Zakir Durumeric. 2023. Hate Raids on Twitch: Echoes of the Past, New Modalities, and Implications for Platform Governance. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 133 (apr 2023), 28 pages. doi:10.1145/3579609

[39] Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. 2023. "Honestly, I Think TikTok has a Vendetta Against Black Creators": Understanding Black Content Creator Experiences on TikTok. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 320 (oct 2023), 31 pages. doi:10.1145/3610169

[40] Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. *ArXiv* abs/2110.00521 (2021). https://api.semanticscholar.org/CorpusID:238253456

[41] Thomas Hehir. 2007. Confronting Ableism. *Educational Leadership* (01 2007).

[42] Sharon Heung, Lucy Jiang, Shiri Azenkot, and Aditya Vashistha. 2024. "Vulnerable, Victimized, and Objectified": Understanding Ableist Hate and Harassment Experienced by Disabled Content Creators on Social Media. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 744, 19 pages. doi:10.1145/3613904.3641949

[43] Sharon Heung, Mahika Phutane, Shiri Azenkot, Megh Marathe, and Aditya Vashistha. 2022. Nothing Micro About It: Examining Ableist Microaggressions on Social Media. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 27, 14 pages. doi:10.1145/3517428.3544801

[44] Tad Hirsch. 2020. Practicing Without a License: Design Research as Psychotherapy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3313831.3376750

[45] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X. Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 205, 21 pages. doi:10.1145/3491102.3517505

[46] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (mar 2018), 33 pages. doi:10.1145/3185593

[47] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 289 (oct 2023), 33 pages. doi:10.1145/3610080

[48] Shagun Jhaver and Amy X. Zhang. 2023. Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society* 0, 0 (2023), 14614448231217993. doi:10.1177/14614448231217993 arXiv:https://doi.org/10.1177/14614448231217993

[49] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. A Trade-off-centered Framework of Content Moderation. *ACM Trans. Comput.-Hum. Interact.* 30, 1, Article 3 (mar 2023), 34 pages. doi:10.1145/3534929

[50] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLOS ONE* 16, 8 (August 2021), 1–22. doi:10.1371/journal.pone.0256

[51] Jigsaw. 2019. Tune: Control the comments you see. https://medium.com/jigsaw/tune-control-the-comments-you-see-b10cc807a171 [Accessed 11-09-2024].

[52] Mark R. Johnson. 2019. Inclusion and exclusion in the digital economy: disability and mental health as a live streamer on Twitch.tv. *Information, Communication & Society* 22, 4 (March 2019), 506–520. doi:10.1080/1369118X.2018.1476575 Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2018.1476575.

[53] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 305 (oct 2021), 44 pages. doi:10.1145/3476046

[54] Sukhnidh Kaur, Manohar Swaminathan, Kalika Bali, and Aditya Vashistha. 2024. Challenges to Online Disability Rights Advocacy in India. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 397, 15 pages. doi:10.1145/3613904.3642737

[55] R.M. Keller and Corinne Galgay. 2010. Microaggressions experienced by people with disabilities in US society. *Microaggressions and marginality: Manifestation, dynamics, and impact* (01 2010), 241–268.

[56] Min Kyung Lee and Katherine Rich. 2021. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 138, 14 pages. doi:10.1145/3411764.3445570

[57] T. A. Lewis. 2022. *Working Definition of Ableism - January 2022 Update.* https://www.talilalewis.com/blog/workingdefinition-of-ableism-january-2022-update Accessed: 2024-12-03.

[58] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "HOT" ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media. *ACM Trans. Web* 18, 2, Article 30 (mar 2024), 36 pages. doi:10.1145/3643829

[59] Yao Lyu and John M. Carroll. 2024. "Because Some Sighted People, They Don't Know What the Heck You're Talking About:" A Study of Blind Tokers' Infrastructuring Work to Build Independence. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 20 (apr 2024), 30 pages. doi:10.1145/3637297

[60] Renkai Ma and Yubo Kou. 2023. "Defaulting to boilerplate answers, they didn't engage in a genuine conversation": Dimensions of Transparency Design in Creator Moderation. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 44 (apr 2023), 26 pages. doi:10.1145/3579477

[61] Abdurahman Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel. 2024. The Virality of Hate Speech on Social Media. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 186 (apr 2024), 22 pages. doi:10.1145/3641025

[62] Kate Manne. 2015. Why I Use Trigger Warnings. *The New York Times* (2015). https://www.nytimes.com/2015/09/20/opinion/sunday/why-i-use-trigger-warnings.html

[63] A McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Cranefield. 2016. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. (1 2016). doi:10.26686/wgtn.12910085.v1

[64] Joy Ming, Sharon Heung, Shiri Azenkot, and Aditya Vashistha. 2021. Accept or Address? Researchers' Perspectives on Response Bias in Accessibility Research. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 20, 13 pages. doi:10.1145/3441852.3471216

[65] Terrance Mok, Anthony Tang, Adam McCrimmon, and Lora Oehlberg. 2023. Experiences of Autistic Twitch Livestreamers: "I Have Made Easily the Most Meaningful and Impactful Relationships". In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 41, 15 pages. doi:10.1145/3597638.3608416

[66] Nicolaas B Moolenijzer and Kristin Dew. 2023. "They know that it works because we are looking for ourselves" – LGBTQ+ TikTok Users' Perceptions and Experiences of Queerbaiting. In *Proceedings of the 25th International Conference on Mobile Human-Computer Interaction* (Athens, Greece) *(MobileHCI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, Article 20, 6 pages. doi:10.1145/3565066.3608705

[67] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. 2022. Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 240, 17 pages. doi:10.1145/3491102.3517608

[68] Richard Nieva. 2024. Here's How Facebook Uses Artificial Intelligence to Take Down Abusive Posts. https://www.cnet.com/tech/tech-industry/heres-how-facebook-uses-artificial-intelligence-to-take-down-abusive-posts-f8/. Accessed: September 10, 2024.

[69] Shuo Niu, Li Liu, and Yali Bian. 2024. Please Understand My Disability: An Analysis of YouTubers' Discourse on Disability Challenges. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 407 (Nov. 2024), 25 pages. doi:10.1145/3686946

[70] Thiago Oliva. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* (04 2021). doi:10.1007/s12119-020-09790-w

[71] 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 2016 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '16)*. Association for Computing Machinery, New York, NY, USA, 369–374. doi:10.1145/2957276.2957297

[72] Jon Porter. 2021. Today I learned about Intel's AI sliders that filter online gaming abuse. https://www.theverge.com/2021/4/8/22373290/intel-bleep-ai-powered-abuse-toxicity-gaming-filters

[73] Jessiva Sage Rauchberg. 2022. *#Shadowbanned: Queer, Trans, and Disabled Creator Responses to Algorithmic Oppression on TikTok* (1st edition ed.). 196 – 210 pages.

[74] Kathryn E. Ringland. 2019. "Autsome": Fostering an Autistic Identity in an Online Minecraft Community for Youth with Autism. *Information in Contemporary Society : 14th International Conference, iConference 2019, Washington, DC, USA, March 31-April 3, 2019, Proceedings. iConference (Conference) (14th : 2019 : Washington, D.C.)* 11420 (April 2019), 132–143. doi:10.1007/978-3-030-15742-5_12

[75] Kathryn E. Ringland et al. 2019. Understanding Mental Ill-health as Psychosocial Disability: Implications for Assistive Technology. In *Proceedings of the Annual ACM Conference on Assistive Technologies (ASSETS)*, Vol. 2019. ACM, 156–170. doi:10.1145/3308561.3353785

[76] Ethan Z. Rong, Mo Morgana Zhou, Zhicong Lu, and Mingming Fan. 2022. "It Feels Like Being Locked in A Cage": Understanding Blind or Low Vision Streamers' Perceptions of Content Curation Algorithms. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 571–585. doi:10.1145/3532106.3533514

[77] Jim Rudd, Ken Stern, and Scott Isensee. 1996. Low vs. high-fidelity prototyping debate. *Interactions* 3, 1 (jan 1996), 76–85. doi:10.1145/223500.223514

[78] Patrawat Samermit, Anna Turner, Patrick Gage Kelley, Tara Matthews, Vanessia Wu, Sunny Consolvo, and Kurt Thomas. 2023. "Millions of people are watching you": Understanding the Digital-Safety Needs and Practices of Creators. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 5629–5645. https://www.usenix.org/conference/usenixsecurity23/presentation/samermit

[79] Shruti Sannon, Jordyn Young, Erica Shusas, and Andrea Forte. 2023. Disability Activism on Social Media: Sociotechnical Challenges in the Pursuit of Visibility. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York,

NY, USA, Article 672, 15 pages. doi:10.1145/3544548.3581333

[80] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1668–1678. doi:10.18653/v1/P19-1163

[81] Joseph S. Schafer, Kate Starbird, and Daniela K. Rosner. 2023. Participatory Design and Power in Misinformation, Disinformation, and Online Hate Research. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 1724–1739. doi:10.1145/3563657.3596119

[82] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 155 (nov 2018), 27 pages. doi:10.1145/3274424

[83] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 368 (oct 2021), 33 pages. doi:10.1145/3479512

[84] Sarita Schoenebeck, Amna Batool, Giang Do, Sylvia Darling, Gabriel Grill, Daricia Wilkinson, Mehtab Khan, Kentaro Toyama, and Louise Ashwell. 2023. Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 485, 16 pages. doi:10.1145/3544548.3581020

[85] Carol F Scott, Gabriela Marcu, Riana Elyse Anderson, Mark W Newman, and Sarita Schoenebeck. 2023. Trauma-Informed Social Media: Towards Solutions for Reducing and Healing Online Harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 341, 20 pages. doi:10.1145/3544548.3581512

[86] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 391, 18 pages. doi:10.1145/3544548.3581538

[87] Ellen Simpson, Samantha Dalal, and Bryan Semaan. 2023. "Hey, Can You Add Captions?": The Critical Infrastructuring Practices of Neurodiverse People on TikTok. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 57 (apr 2023), 27 pages. doi:10.1145/3579490

[88] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini (Eds.). 2021. *SoK: Hate, Harassment, and the Changing Landscape of Online Abuse.*

[89] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein (Eds.). 2022. *"It's common and a part of being a content creator": Understanding How Creators Experience and Cope with Hate and Harassment Online.*

[90] Alexandra To, Hillary Carey, Geoff Kaufman, and Jessica Hammer. 2021. Reducing Uncertainty and Offering Comfort: Designing Technology for Coping with Interpersonal Racism. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 398, 17 pages. doi:10.1145/3411764.3445590

[91] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2023. Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models. *CoRR* abs/2307.09209 (2023). doi:10.48550/ARXIV.2307.09209 arXiv:2307.09209

[92] Jayne Wallace, John McCarthy, Peter C. Wright, and Patrick Olivier. 2013. Making design probes work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3441–3450. doi:10.1145/2470654.2466473

[93] Yihe Wang and Kathryn E. Ringland. 2023. Weaving Autistic Voices on TikTok: Utilizing Co-Hashtag Networks for Netnography. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 254–258. doi:10.1145/3584931.3606995

[94] Mark Warner, Angelika Strohmayer, Matthew Higgs, and Lynne Coventry. 2024. A Critical Reflection on the Use of Toxicity Detection Algorithms in Proactive Content Moderation Systems. arXiv:2401.10629 [cs.HC] https://arxiv.org/abs/2401.10629

[95] Gregor Wolbring. 2008. The Politics of Ableism. *Development* 51, 2 (2008), 252–258. doi:10.1057/dev.2008.17

[96] Lana Zhang and Ravisha SK. 2023. Flag Harmful Content: Using Amazon Comprehend for Toxicity Detection. https://aws.amazon.com/blogs/machine-learning/flag-harmful-content-using-amazon-comprehend-toxicity-detection/. Accessed: September 10, 2024.