"I thought it was my mistake, but it's really the design": A Critical Examination of the Accessibility of User-Enacted Moderation Tools on Facebook and X

SUDHAMSHU HOSAMANE, Rutgers University, USA
ALYVIA WALTERS, University of Kentucky, USA
YAO LYU, University of Michigan, USA
SHAGUN JHAVER, Rutgers University, USA

As social media platforms increasingly promote the use of user-enacted moderation tools (e.g., reporting, blocking, content filters) to address online harms, it becomes crucial that these controls be accessible to everyone. We evaluate the accessibility of moderation tools on two mainstream platforms – Facebook and X – through interviews and task-based walkthroughs with 15 individuals with vision impairments. Adapting insights from public administration scholarship, we introduce the lens of *Administrative Burden of safety work* and foreground three interleaved costs that users with vision loss incur while interacting with moderation tools: *learning costs* (understanding what controls do and where they live), *compliance costs* (executing multi-step procedures under screen reader and low-vision conditions), and *psychological costs* (experiencing uncertainty, stress, and diminished agency). Our analysis bridges the fields of content moderation and accessibility in HCI research and contributes (1) a cross-platform catalog of accessibility and usability breakdowns affecting safety tools; and (2) design recommendations for reducing this burden.

Additional Key Words and Phrases: Content Moderation, Blind, Low Vision, Usability

ACM Reference Format:
Sudhamshu Hosamane, Alyvia Walters, Yao Lyu, and Shagun Jhaver. 2026. "I thought it was my mistake, but it's really the design": A Critical Examination of the Accessibility of User-Enacted Moderation Tools on Facebook and X. *Proc. ACM Hum.-Comput. Interact.* X, CSCW1 (October 2026), 38 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Online safety is a growing problem, encompassing challenges of persistent harassment and hate, fast-spreading misinformation, technology-facilitated privacy and sexual harms, and deepfake-driven impersonation [13, 25, 39, 72, 93, 94, 99]. For marginalized populations, especially, online safety is an increasingly concerning topic [18, 34, 37, 61]. For instance, prior research on people with visual impairments has documented the online safety problems they encounter when using social media, including targeted attacks by trolls and harassment; these problems are exacerbated by accessibility barriers and inadequate response from platform management [40, 54, 57, 58].

Social media platforms rely upon content moderation infrastructures for online safety. Traditionally, platforms used third-party [69] and volunteer moderators [60, 62, 102] as well as automated

Authors' addresses: Sudhamshu Hosamane, sudhamshu.hosamane@rutgers.edu, Rutgers University, New Brunswick, NJ, USA; Alyvia Walters, alyvia.walters@rutgers.edu, University of Kentucky, New Brunswick, NJ, USA; Yao Lyu, yaolyu@umich.edu, University of Michigan, Ann Arbor, MI, USA; Shagun Jhaver, shagun.jhaver@rutgers.edu, Rutgers University, New Brunswick, NJ, USA.

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 ACM 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.
© 2026 Association for Computing Machinery.
2573-0142/2026/10-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

tools [11, 45, 51] to perform "safety work," defined in this study as the various types of labor involved in limiting the reach and presence of harmful content online through the deployment of content moderation tools. Recently, however, platforms are increasingly delegating safety work to end users themselves through *user-enacted moderation* tools such as reporting, blocking, and content filtering [15, 27, 46, 48]. For example, recent announcements by Meta[1] signal greater dependence on user-submitted reports, leaving automated systems to prioritize only illegal or highest-severity content. This is a trend across Silicon Valley, which coincides with the U.S. political shift toward strongly neoliberal orientations: without pressure to stringently moderate speech on their platforms, Big Tech is outsourcing safety work to users rather than contending with the financial burden and regulatory messiness of more sophisticated oversight. Arguably, through deploying U.S.-centric rhetorics of freedom of speech, platforms are claiming that these changes mark a democratic celebration of freedom of expression while actually diminishing the quality and safety of content on their sites [17].

Given these trends, it is vital to advocate for keeping platforms accountable for maintaining minimal site-wide safety standards. Beyond that, it is also crucial that user-enacted moderation tools, which the platforms themselves are pushing for, are usable and accessible to all users. However, prior work in Human–Computer Interaction (HCI) shows that adopting these tools is often costly: users must invest time, cognitive effort, and careful social reasoning to control, configure, and deploy them effectively [15, 47, 48]. For users with visual impairment, this problem can be more serious. A substantial body of HCI and accessibility research shows that even *basic*, single-step interactions on mainstream social media—discovering menus, interpreting icon-only buttons, and reading posts—can be inconsistent or difficult to perform non-visually [50, 63, 70]. Against this backdrop, we are concerned that inaccessible moderation controls could undermine user safety, impose disproportionate interaction burdens [6], and deny people with vision-related disabilities the agency that the use of these tools promises.

Driven by this concern, we examine the accessibility of user-enacted moderation tools through in-depth, task-based interviews with 15 social media users with visual impairments on Facebook and X (formerly Twitter). The primary goal is to recognize accessibility barriers through a qualitative approach, so we prioritize discovering accessibility problems rather than assessing or comparing task completion time across participants. To ensure participant privacy and avoid affecting their real-life social media usage, we set up a standardized and realistic environment on Facebook and X, where we provided examples of harmful content using dummy or private accounts. Leveraging these example posts, we asked participants to complete the following tasks: report a hateful post, report a dummy spam group/community, block a user, change post visibility settings, unblock the previously blocked user, and adjust sensitivity settings. Rather than just asking whether these controls are *perceivable* or *usable*, we center *operational accessibility*: whether users with blindness or low vision (BLV) can *complete* safety actions end-to-end.

Our analysis shows how current designs obscure or complicate essential controls across the entire workflow: locating an entry point, navigating multi-step pages with inconsistent layouts, understanding the consequences of actions, and verifying outcomes that are often hidden. We frame these breakdowns using the lens of *Administrative Burden* —a public administration frame [64] which allows us to distinguish *learning*, *compliance*, and *psychological* costs—to articulate how small frictions at any stage compound and propagate, making later steps harder and eroding user confidence.

[1]Changes in policy enforcement to now rely more on user reports: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

This paper extends HCI literature at the intersection of accessibility, online safety, and platform governance. While prior work has examined how blind users experience platform moderation [54] and the accessibility of misinformation warnings [75], little has been done to comprehensively examine users with vision impairments' use of user-enacted moderation tools at the interaction level. We address this gap by providing a detailed evaluation of how users with blindness and low vision navigate multi-step moderation workflows from initial discovery through outcome verification. Our task-based methodology centers on *operational accessibility*—whether users can complete safety actions end-to-end. We frame our findings using the concept of Administrative Burden, which provides a precise vocabulary—learning, compliance, and psychological costs—for articulating how small frictions at each workflow stage compound and propagate, ultimately eroding user confidence and safety. Grounded in these findings, we offer design recommendations to make moderation controls more equitable and navigable.

2 Related Work

We first provide an overview of *content moderation research on social media platforms*, highlighting especially the work on user-enacted moderation tools and an assumed "sighted default" in their design (Section 2.1). We then survey *digital accessibility work*, cataloging the barriers that users who are blind or have low-vision face in routine social media tasks (Section 2.2). Synthesizing these strands reveals an under-examined problem—the inaccessibility of safety tools for users with visual disabilities—that our study directly tackles.

2.1 Overview of User-Enacted Content Moderation on Social Media

Social media platforms bear the key responsibility of shaping online safety by defining and enforcing community standards of appropriate conduct [38, 68]. In enforcing these standards, they identify and regulate a wide array of problematic content, from illicit material and coordinated harmful behavior to hate speech, misinformation, and interpersonal harassment [73]. Platform regulation is critical given that such norm violations can profoundly deteriorate user well-being, public discourse, and the very integrity of these digital environments [28, 65]. Most efforts to address these issues have relied on centralized, platform-driven interventions. These typically involve a dynamic interplay between increasingly sophisticated automated detection systems designed to flag inappropriate content at scale and vast teams of professional human moderators tasked with adjudicating policy violations, especially those that are too nuanced or unclear to be judged by automated systems [28, 69]. While this mixed initiative approach is necessary to handle the large volume of content, it presents problems such as the potential for algorithmic bias and the significant, well-documented psychological burden placed upon human reviewers who are incessantly exposed to distressing material [60, 69, 82].

Recognizing the limitations of purely centralized models, platforms are increasingly delegating the responsibility of safety work to end-users. This includes community-led moderation (e.g., by end-users who serve as volunteer moderators on Reddit or in Facebook Groups) [60, 74] and, more pervasively, the provision of individualized user-driven moderation tools [48]. Features like blocking, muting, and reporting content are now standard across platforms, aiming to empower individuals to tailor their experiences and flag harmful content for platform review [15, 28, 47]. Scholarly inquiry into user-enacted moderation has explored a variety of dimensions from the perspective of the general user base, including the motivations behind using these tools (i.e., why users block [47] or report [106] others), the perceived fairness and effectiveness of reporting systems [76], and user preferences regarding the use of personalized content filters versus broader platform-level interventions to address norm violations [48].

Most closely related to our work, Heung et al. [41] have explored how configurable AI-assisted filters can protect users with disability from ableist hate, highlighting the need for granular, user-controlled safety mechanisms. As platforms delegate more safety work to users, the value of such control is determined by whether the people they are meant to protect can reasonably exercise it. Our study examines this question through experiences of users with BLV finding and using existing moderation tools, and attending to the difficulties involved across completing various safety tasks. This connects the promise of user-enacted moderation to the accessibility of everyday social-media interfaces, which we discuss next.

2.2 Accessibility in Social Media: Persistent Barriers for Users with Visual Disabilities

Social media platforms like Facebook and X are integral to contemporary life for many, including those with vision-related disabilities [30, 32, 53]. These users were early adopters, benefiting from the platforms' initial text-based nature [95, 103], and continue to engage heavily with them [29]. Yet widespread accessibility barriers persist—not by accident, but due to systemic neglect of core user-centered design principles [67, 101]. The Web Content Accessibility Guidelines (WCAG) are a World Wide Web Consortium standard for making web content accessible to people with disabilities, specifying testable requirements organized under four principles: Perceivable, Operable, Understandable, and Robust (POUR) [1]. Failures to meet these requirements, such as insufficient contrast or missing text alternatives, can create accessibility barriers for users with visual impairments.

The most fundamental barriers to accessibility on social media usually undermine the *"Perceivable"* and *"Operable"* principles of WCAG. Visual content such as images and memes often lack appropriate alternative text, rendering them incomprehensible to screen readers [31]. Automated captioning systems exist, but their accuracy and context awareness remain limited, so they have not fully solved this problem [81]. Critically, interactive elements like buttons and menus often lack proper semantic labels. This forces screen reader users into laborious trial-and-error navigation for basic tasks [63], as they often cannot perceive the function of a control or operate the interface effectively.

Users with low-vision face a different, yet equally challenging, set of barriers. While they may use aids like virtual screen magnifiers, interfaces are often not *"Robust"* enough to handle them gracefully. Magnification can disrupt page layouts, requiring tedious panning that obscures page context and violates basic usability tenets [84, 85]. Further, when accessibility-based gestures conflict with app-based gestures or become confusing to track, it makes routine actions cognitively taxing and error-prone [3].

These persistent accessibility gaps in everyday social media use suggest a grim prognosis for the usability of a broader range of tools. When basic tasks like reading a feed, following links, and manipulating buttons remain difficult for users with vision loss [55, 56], it is reasonable to expect that more complex, multi-step features such as reporting flows or other trust-and-safety tools will be even less accommodating. Independent accessibility audits support this concern: investigators have documented unlabeled controls and other flaws that make moderation workflows effectively unusable with screen readers [88]. While such audits begin to offer valuable initial insights into these challenges, they were brief spot-checks. As such, they stopped at listing missing labels—without tracking participants across multi-screen reporting process, measuring task-completion drop-offs, or examining other subtler obstacles that users with vision impairment could face. Broader analyses of accessibility research trends suggest that nuanced investigations into complex, safety-critical workflows like content moderation for users with vision-related disabilities are still relatively scarce compared to studies on more foundational interaction or consumption tasks [43].

This recurring pattern of inaccessibility in safety-critical tools extends beyond isolated usability flaws; it reflects a systemic failure to apply principles of design justice and critical disability

studies [14, 80, 83]. This oversight effectively violates the disability rights imperative of "*Nothing About Us Without Us*" [12]. When safety tools are not designed inclusively, users with vision-related disabilities are forced to perform significant "*invisible adaptation work*" just to achieve basic safety [6], a burden likely intensified for those who may already face disproportionate online harm [40, 42].

Emerging studies are beginning to highlight the tangible, severe consequences of this accessibility gap in safety-critical contexts. Lyu et al. [54] found that TikTok users who had blindness struggled to report harassment due to inaccessible interface elements, and some ended up getting unfairly sanctioned as a result. Similarly, Sharevski and Zeidieh [75] revealed that misinformation warnings on YouTube, Facebook, and TikTok were effectively invisible to users with visual impairments, leaving them unprotected from harmful content. These studies demonstrate that accessibility failures in moderation tools compound vulnerabilities—users cannot protect themselves and may face punitive consequences for their attempts.

Prior research has extensively documented the adaptation work disabled users perform when navigating inaccessible systems, including workarounds [6], emotional labor [16], ongoing accommodations [77, 98], and infrastructuring work to build independence in hostile digital environments [56]. These frameworks powerfully articulate the *existence*, *scale*, and *inequity* of extra work disabled users perform. However, they were not designed to analyze the *temporal and procedural dynamics* of multi-step workflows that require users to navigate platform policies—such as selecting violation categories when reporting harassment or configuring privacy controls. These are not simple one-click actions; they are structured procedures where users must interpret policy categories, make choices across multiple screens, and often receive little feedback about outcomes.

Our study complements this literature by adopting the administrative burden framework from public administration [64], which examines people's experiences of navigating bureaucratic procedures. This framework distinguishes *learning costs*—discovering programs, understanding eligibility and benefits, and learning how to obtain and maintain access; *compliance costs*—the time and effort required for forms, documentation, waiting, and travel; and *psychological costs*—stress, frustration, stigma, and loss of autonomy that accompany these encounters. Researchers have used this framework to examine how institutional procedures and technologies shape access to support. Yates et al. [105] showed how administrative burdens hindered women's access to sufficient support through Australia's National Disability Insurance Scheme. In LLM-assisted access to public benefits, Jo et al. [49] identified potential reductions in existing burdens alongside new costs of learning to interact with the system, verifying its advice, and concerns about losing human advocacy. We adapt this typology to platform-administered safety workflows, treating interactions with protection controls as service encounters. This framework is well-suited for analyzing platform safety tools because content moderation workflows function as governance systems where users interact with platform policies through structured, multi-step procedures. The framework therefore helps us locate where interface design creates or relieves burdens for non-visual interaction and trace how these costs *accumulate and cascade* across procedural steps (Section 4).

Applying this lens, we examine how users with vision impairments experience user-enacted moderation tools on Facebook and X, documenting the accessibility barriers they encounter and the administrative costs these barriers impose. This investigation is guided by the following research questions:

(1) **RQ1:** How do users with blindness and low-vision currently perceive, navigate, and attempt to utilize common user-enacted moderation tools (e.g., reporting, blocking, content filtering) available on mainstream social media platforms?

(2) **RQ2:** What specific usability and accessibility barriers do users with vision impairment encounter when engaging with these moderation interfaces, and how do these barriers impact their ability to manage their safety and online experience effectively?
(3) **RQ3:** What design improvements and systemic changes are needed to create genuinely accessible and equitable user-enacted moderation tools for people with visual impairments?

3 Methods

3.1 Study Participants and Protocol

This study employed a qualitative methodology to investigate the experiences of individuals with blindness or low-vision with user-enacted content moderation tools. We studied tools on two major social media platforms: Facebook and X. These platforms are among the most widely used by people with visual impairments [5, 29]. Their history of accessibility research [32, 104] and earlier investment in dedicated accessibility teams [23, 90] also motivated us to examine the accessibility of user-enacted moderation tools on these platforms. Our approach centered on observing participants as they performed a series of representative moderation, privacy, and safety-related tasks on their own mobile devices. Ethical approval for all study procedures was obtained from the Institutional Review Boards (IRB) at our universities.[2]

We recruited fifteen participants from a US state[3] who were 18+ years old, identified as having BLV, were regular Facebook or X users, were proficient with assistive technology, and were fluent in English. Recruitment proved challenging though ultimately successful; after unproductive online advertising and unsuccessful outreach to local organizations and universities, participants were sourced through a community mailing list and subsequent word-of-mouth referrals.

Data collection occurred between November 2024 and April 2025, with the majority of interviews (13 out of 15) conducted between February and April 2025. To enhance ecological validity and participant comfort, all study sessions were conducted in person and at locations chosen by the participants, which included their homes, local libraries, or quiet coffee shops. Each participant used their own smartphones,[4] their preferred screen reader software (all 15 participants in this cohort used VoiceOver on iOS), and their preferred social media platform (six used X, the rest opted for Facebook). This ensured that interactions took place within a familiar technological setup. Sessions were comprehensive, lasting between 2 hours and 5 minutes and 3 hours and 15 minutes (median duration: approximately 2 hours and 15 minutes), allowing sufficient time for rapport-building, task completion, and a debriefing at the end. Participants were encouraged to take adequate breaks in between tasks and were compensated with a $50 Amazon gift card for their time, provided at the conclusion of each session irrespective of task completion. We also helped them load the gift card onto their accounts.

A detailed summary of participant demographics, including pseudonyms, age, social media platform used during the study, self-reported visual condition, and years of experience with VoiceOver and the respective platform, is provided in Table 1. The cohort ranged in age from 21 to 81, with most having extensive experience (median of approximately a decade) using screen readers and mobile technology. Thirteen participants reported having complete blindness or having only minimal light perception, while the remaining two had severe low vision. It is crucial to underscore, however, that the lived realities of these individuals are far more nuanced than a table can fully capture. Participants experienced vision loss at different stages of life—some were blind since birth, others

[2]We will disclose the university name after the peer review process completes.
[3]We will disclose the location after the peer review process completes.
[4]All participants owned an iPhone and used it as their primary device for performing moderation tasks during the interview. They were given the option to use any digital device they were comfortable with.

Table 1. Participant Demographics and Media Usage Background

Pseudonym	Age	Platform	Condition	VoiceOver (yrs)	Platform (yrs)
Shreya	21	X	High myopia + retinal issues	5	13
Kiara	25	X	Leber Congenital Amaurosis	12	12
Samuel	26	Facebook	Retinal detachment (lights only)	10+	14
Jamila	30	Facebook	Congenital glaucoma (light only)	14	15
Maya	35	X	Pseudotumor Cerebri (light only)	6	13
Adriana	36	Facebook	Retinitis Pigmentosa (RP) (legally blind)	15	16
Jessica	36	Facebook	Congenital glaucoma (total blind)	11	13
Grant	38	X	Glaucoma + Aniridia (light only)	9	5
Mateo	45	X	Glaucoma (light perception in one eye)	4.5	14
Cristian	51	X	Glaucoma (totally blind since 17)	13	6
Frank	57	Facebook	RP + Stargardt (light only)	6	15
Nicole	62	Facebook	RP (legally blind, tunnel vision)	<1	15
Laverne	63	Facebook	Glaucoma + Failed Corneal Transplant	12	1
Vasu	66	Facebook	Glaucoma (total blind)	15	13
Patricia	81	Facebook	Optic Neuritis, RP (total blind)	13	15

experienced significant vision loss later in life (e.g., after age 60), and some were in the ongoing process of losing their sight. Consequently, their journeys with technology varied: some learned to use social media for the first time after vision loss, while others had to relearn interaction paradigms using a screen reader after years of sighted use.

3.2 Task Design and Administration

To investigate how users with visual disabilities engage with content-moderation controls—and the barriers they may encounter—we designed a sequence of *six* tasks based on the content-moderation mechanisms available on Facebook and X (discussed in Section 2.1). These tasks were grounded in platform governance research on user-enacted moderation. Grimmelmann's taxonomy of online moderation identifies *exclusion* (removing users or content from one's experience) as a fundamental moderation verb that platforms delegate to users through tools such as blocking, muting, and filtering [33]. Crawford and Gillespie argue that reporting tools operate as a "practical and symbolic linchpin" in platform governance, even as their operation remains poorly understood from users' perspectives [15]. Jhaver et al. further document how blocking tools and blocklists function as harassment-prevention mechanisms, yet often fail to provide adequate protection in practice [47]. Building on this literature and the set of controls present on both platforms, our six tasks span three primary categories of user-enacted moderation: *reporting* (flagging content or accounts for platform review), *interpersonal exclusion* (blocking and unblocking), and *content filtering* (adjusting visibility and sensitivity preferences). The task sequence progressed from basic content interaction and reporting to more targeted safety, privacy, and preference-setting actions, reflecting common user-directed moderation workflows.

For each task, we evaluated accessibility against established interaction-design principles that have been widely applied in accessibility research [52, 70]. Specifically, we assessed *discoverability* (can users locate the feature's entry point?), *navigability* (can users complete the multi-step workflow?), *feedback* (does the interface communicate progress and outcomes?), *consistency* (do similar features behave similarly across contexts?), *user control* (can users customize their experience?), and *reversibility* (can users undo actions?) [66, 67, 78]. These principles align with WCAG success criteria related to operability and understandability [1, 96].

3.2.1. *Task Environment and Stimuli*

To ensure a standardized yet realistic context for task performance while prioritizing participant privacy and avoiding unintended consequences for their personal social media accounts, all tasks were conducted using controlled account environments.

For interactions on X, two study-specific dummy accounts were created by the research team: 'Jenny Chang' and 'Marcus Chang'.[5] The 'Marcus Chang' account was used to post simulated problematic content directed at 'Jenny Chang's' profile by tagging the latter's username. No content was drawn from real posts or real accounts. To avoid exposing participants to distressing stimuli, this content often included placeholders rather than displaying actually offensive material. Further, we arrived at the decision to include placeholder text because what "counts" as hateful or offensive varies from person to person. We reasoned that placeholder text would elicit a less constrained approach to the moderation task across participants, without introducing confusion on what type of content qualifies as hateful or offensive. We asked participants to read each placeholder as standing for whatever comment they themselves would find hateful.

Each post paired a plainly worded message with a placeholder naming the harm it stood for, so that the context made the attack legible while the offensive material itself was withheld. For example, in Figure 1, "Hey [name], are you still wearing that rag? [This message contains a targeted attack on your clothing style]", and on X (Figure 5) "You are annoying [personal attacks, slurs here], keep to yourself". We provide more examples in the Appendix (Figure 7).

During these sessions, we helped participants login into the dummy account of 'Jenny Chang', and instructed them: "Imagine you are Jenny Chang, and you've encountered these posts from Marcus Chang on your profile. If you were in her shoes, how would you proceed doing x (the tasks to follow)?" We followed an equivalent procedure on Facebook.[6] We then explicitly specified the action to take, such as reporting the post or blocking its author, without divulging how to do so. Our goal was to study the end-to-end operability of these security tools. We provided assistance if a participant gave up, explicitly requested help, or could not complete the task within 15 minutes. We provide the complete interview protocol in the Appendix.

3.2.2. *Moderation Tasks and Rationale*

We asked participants to complete the following tasks:

(1) **Report a Hateful Post:** Participants encountered a simulated post containing hostile or discriminatory content and identified the platform's flagging tool in order to report the post. We designed this task to assess the initial *discoverability* [67] and *navigability* of flagging procedures through different context menus until they completed the task.
(2) **Report a Dummy Spam Group/Community:** We asked participants to navigate to a pre-created, inactive fake group (Facebook) or community (X) and report it for spam. This task broadened the scope from individual posts to a group, enabling us to evaluate different reporting contexts for *consistency* [78].
(3) **Block a User:** We instructed participants to block the account that was posting problematic content. This simulation provided a direct response to harassment and allowed us to evaluate the process of enacting interpersonal restrictions, a key aspect of *user control and freedom* [66].

[5]These pseudonyms were randomly generated and do not represent real individuals; realistic-sounding names were chosen to ensure we weren't blocked from the platform.

[6]Facebook's strict scrutiny of fake accounts prevented creation of dummy profiles, so we used the first author's rarely used personal account (with a limited social graph) as the recipient and a collaborator's low-activity account as the sender. The collaborator posted placeholder content that tagged the author; visibility of the posted messages was restricted to just the two accounts, mirroring the X setup while safeguarding participant privacy.

(4) **Change Post Visibility Settings:** We asked participants to modify the audience of a sample post (i.e., from public to friends-only). This task explored participants' ability to manage content privacy, reflecting *user control* over their own information dissemination.
(5) **Unblock the Same User:** After completing the account-blocking task, we asked participants at a later stage to locate and reverse the blocking action. This tested the *reversibility* of actions [67] and *ease of redressal*, crucial for managing dynamic social boundaries.
(6) **Adjust Sensitivity Settings:** Both Facebook and X offer content filtering (e.g., for political or sensitive content); we asked participants to locate and modify these preferences to see 'less' of this content (from the previously set 'default' level). This task allowed us to examine the accessibility of controls related to *personalized content experiences* and *proactive filtering*.

3.2.3. Think-Aloud Protocol and Data Capture

Throughout the task-based portion of the session, we employed a combination of *concurrent* and *retrospective* think-aloud protocols [71]. Participants were encouraged to verbalize their thoughts, actions, and expectations as they interacted with the interfaces. We accounted for expert screen-reader users' tendency to listen at very high speech rates and to process information rapidly. This intense focus can sometimes lead to participants forgetting to verbalize their thought processes concurrently. Therefore, to ensure rich data capture, interviewers provided gentle, periodic nudges such as "What is VoiceOver saying now?" or "What are you trying to do on this screen?" to encourage continuous narration, especially during extended silences.

Following each task, or at points of significant struggle, we asked brief retrospective questions (e.g., "What were you searching for there?" or "Can you try to remember what you found difficult about that step?") to clarify participants' mental models and strategies. We phrased questions in simple language and provided ample time for task completion and screen reader use. The interviewer minimized direct observation of the screen when it might disrupt the participant's auditory focus, relying primarily on the verbalized think-aloud data. All sessions were audio-recorded and later transcribed verbatim for analysis.

3.3 Data Analysis

We employed reflexive thematic analysis following Braun and Clarke's guidelines [7–9], foregrounding the researcher's interpretive role in developing themes. The first author transcribed all audio recordings using AssemblyAI,[7] and verified portions of each transcript for correctness.

Our research questions and task design (Section 3.2.2) guided our attention to how participants found, understood, and operated moderation controls. Within this focus, we examined participants' actions and accounts of their assumptions, expectations, difficulties, and resolution strategies. The first author conducted iterative, line-by-line coding of every transcript, generating 155 initial codes. We maintained a working codebook to document the developing codes, updating it as coding progressed. Our coding was grounded in a constructionist epistemology, i.e., we actively produced meaning through engaging with participants and their social contexts, and mapped the diverse ways that participants constructed their moderation experiences. We prioritized semantic codes over latent codes. Earlier transcripts were revisited and coding was revised as our interpretations developed. Weekly co-author discussions examined the working codebook, candidate themes, and analytic memos to refine our interpretations.

Through iterative coding and discussion, we organized the codes into five broad analytical themes: *Discoverability* (locating entry points), *Navigability* (moving through multi-step workflows till the completion of task), *Adaptation and Agency* (workarounds, control, and customization),

[7]https://www.assemblyai.com

Cognitive Load (mental effort, memory, and attention), and *Expectation Mismatches* (gaps between user assumptions and interface behavior). These themes provided a starting point for interpreting relationships across participants' experiences as we show in the *Findings* section.

Following coding, we saw parallels between administrative burden [64] and the aforementioned themes. We experimented using that as an interpretive lens to examine how learning, compliance, and psychological costs accumulated across these experiences. This lens informed the organization of our findings into three composite categories: *Discovering and Navigating Safety Controls* (Discoverability, Navigability, and related burdens); *Uncertainty in User Controls* (Mismatch in user expectations or assumptions regarding the behavior of an interface or its affordance and the resulting issues); and *Adaptation, Emotional Labor, and Perceived Agency* (User adaptations and experiences around agency to complete the safety task). We used Excel for code matrices and NVivo for transcript annotation.

3.4 Researcher Positionality

All members of the research team are sighted, and the first author led the interview and analysis work with the guidance of the other three. We acknowledge that this positionality influences our approach: for instance, we may not immediately notice accessibility issues that users with BLV experience daily. To address this, our team included advisors with expertise in BLV accessibility and content-moderation systems. We also consulted informally with other colleagues who work on accessibility and disability justice to review our procedures and interpretations. The second and fourth authors are specialists in user-centered social media governance research, while the third author is an expert in accessible technology evaluation; their prior experiences shaped our study design. We approached our analysis reflexively, maintaining an active commitment to disability-centered design principles by prioritizing participants' own descriptions and framing of their experiences and actively questioning our interpretations to ensure they faithfully represented the participants' perspectives.

4 Findings

Our findings reveal a pervasive pattern of accumulated design shortcomings within Facebook and X's moderation interfaces that impose a significant burden on users with vision loss. We frame these results as the *administrative burden of safety work*—the *learning*, *compliance*, and *psychological* costs people incur when they try to protect themselves online (as defined in Section 2.2).

Participants reported varying levels of familiarity with moderation controls. **Eight** of the 15 participants reported no prior use of reporting on their study platform, although some had used other controls, such as muting or audience settings, and only **two** disclosed prior use of the blocking or reporting. Awareness of a feature did not necessarily entail knowing how to use it: Laverne had heard about blocking and privacy controls from her sister but did not know how to operate them, while Shreya described herself as only "somewhat" familiar despite prior use. Participants thus brought different levels of prior experience to individual moderation tasks.

We begin with discovery and navigation hurdles (Section 4.1), where learning costs (finding/understanding entry points) and compliance costs (locating low-visibility icons, multi-step flows, detours, restarts) accumulate. We then show how design inconsistencies (Section 4.2) undermine participants' expectations and add learning costs (misaligned mental models) and psychological costs (doubt, second-guessing). Finally, we trace consequences (Section 4.3): workarounds and rework that raise compliance costs, alongside emotional labor and diminished agency as psychological costs. Figure 8 in the Appendix shows how these costs combined for one task, reporting a hateful post.

4.1 Discovering and Navigating Safety Controls

Participants usually navigate digital devices with a screen reader like VoiceOver. On each screen, Voiceover typically focuses on the first UI (user interface) element.[8] Interaction relies on a sequence of gestures: linear navigation (swiping left/right to move between elements sequentially), touch exploration (dragging a finger to identify elements by location), and custom actions (e.g., swiping up/down on an element to reveal context menus—akin to a long press). Effective navigation thus depends entirely on audible cues from properly labeled elements (buttons, links, text, etc.) and a predictable interface sequence.

Successfully using any moderation tool—the task we continually asked participants to undertake during our sessions—unfolds in two critical stages for all users, including users with BLV. First, users must *discover* the initial entry point—typically an icon or menu link that houses controls like "Report" or "Block." Second, they must *navigate* the subsequent multi-step workflow of panels and selection screens to complete the action. A failure at either stage can render the entire feature unusable. Accordingly, we structure our findings to follow this user journey, first examining the barriers to discoverability and then the challenges of workflow navigability.

4.1.1. Discoverability Challenges

A recurring theme across our interviews was the difficulty participants faced in *locating the initial entry point*—typically an icon or button—that would allow them to begin navigating the sequence of options required to complete a moderation task. Considering this through the lens of administrative burden, this immediately presented an obvious *learning cost* to participants. One particularly problematic element was the small, often low contrast "•••" menu icon—commonly known as the meatball menu[9]. This icon often houses crucial moderation options such as reporting a post, blocking a user, or flagging a group based on the context it is presented in. Discoverability failures stemmed from two main issues: *counterintuitive focus order* for screen reader users and *poor visual contrast* of menu icons for users with low-vision.

When Jessica attempted to report an offensive Facebook post, she opened it and began methodically exploring its layout for the reporting control. VoiceOver announced other elements within the post, such as the author, time-stamp, and Like/Comment controls, but the "More Options" menu icon, which is the entry point for reporting, was never announced in this expected sequence (Figure 1). *"It didn't speak it to me... I had to go out of the post to get it,"* she explained. Frustrated, she had to navigate back to the main news feed—the list of all posts—where she could access a different context menu on the post summary that finally revealed the report function. The detour was a compliance cost she paid before the reporting flow had begun, and her frustration was a psychological cost.

Prompted by the sighted interviewer, who confirmed the menu icon was visually present at the top-right of the expanded post view, Jessica returned to investigate and found that the focus order was reversed. She found that the "More Options" menu was only accessible with VoiceOver by swiping *left* from the 'Back' button—a counter-intuitive sequence, as screen reader users expect right swipes to move focus forward through a screen's layout, typically from left to right. Vasu encountered the same anomaly, calling it *"backwards from how the screen looks."* This counterintuitive focus order is not unique to Facebook; accessibility research has documented widespread focus and ordering problems in mobile applications [70]. What distinguishes this barrier in the

[8]VoiceOver visually indicates its current focus with a black rectangle around the focused item (as shown in Figures 1a and 1b).

[9]We refer to the ubiquitous meatball menu icon, used on both Facebook and X, to group additional actions, simply as the menu icon henceforth.

(a) VoiceOver focuses on the "Back" button (top-left) by default.

(b) VoiceOver moves focus to the "More options" menu icon (top-right) after swiping left (*not right*) from "Back."

Fig. 1. Misleading focus order in Facebook's interface

safety context is its gatekeeping effect: when the entry point to reporting is effectively hidden in an illogical focus sequence, users cannot even begin the protective action they seek, highlighting the exceedingly high *learning* and *compliance costs* for users with BLV.

Participants with low-vision often knew where moderation controls should be but struggled to locate them. On X, reporting tools are placed in a pale white menu icon that blends into the interface and does not scale with iOS text-size settings. When Shreya—who has severe myopia and uses iOS Display Zoom (to increase text size)—was asked to report a post on X, she tapped the upper-right corner "on muscle memory," yet admitted: *"I can't see the dots unless I hold the phone right up to my face... they're not really visible, it's like a little faded gray, not completely in your face."* Even with Zoom-in enabled, the icon stayed small and washed-out, leading her to remark: *"I think even someone with perfect sight might miss it at first glance."* Locating the icon took her extra time and effort, a *compliance cost*.

Facebook Groups presented a similar visibility problem. On iOS, the horizontal *menu* icon—which contains options like reporting—is rendered directly over the large decorative cover photo (Figure 2). Nicole, a participant with low-vision, attempted to flag a dummy group for spam but could not locate the icon and didn't know where to search for it; its low-contrast design blended into the photo and, as she put it, *"just disappeared,"* presenting a *learning cost* which precluded her ability to engage in safety work. Nicole suggested placing the icon on a contrasting background: *"put a box around the three dots so that whatever you put up here, those three dots are in a box."* Our technical analysis[10] confirmed this was systemic: the icon failed the WCAG 2.1 non-text contrast threshold of 3:1 in five of Facebook's six[11] default cover images. On X, the default menu icon on posts also failed this test, but the platform's built-in "Increase color contrast" option in the accessibility setting rectified the issue[12].

4.1.2. Workflow Navigability Challenges

Once past the initial challenge of discovering the entry point for moderation, participants faced a second set of hurdles related to *navigability*. Whereas *discoverability* involved finding the correct icon or button to initiate a moderation task, navigability here refers to completing the subsequent

[10]To assess potential low-contrast issues, we avoided relying on developer tools like Xcode's Accessibility Inspector, which cannot extract color values from graphical icons. Instead, we disabled True Tone and Night Shift on an iPhone 13, took lossless PNG screenshots, and used two separate color samplers to capture both the icon's foreground color and the surrounding background tones behind it. We used these values to test contrast sufficiency.
[11]Link to the group cover photos we tested for contrast: https://figshare.com/s/6b500311ac94d1b158e3
[12]We provide screenshots for this here: https://figshare.com/s/4969cc7d674f2725d7a6

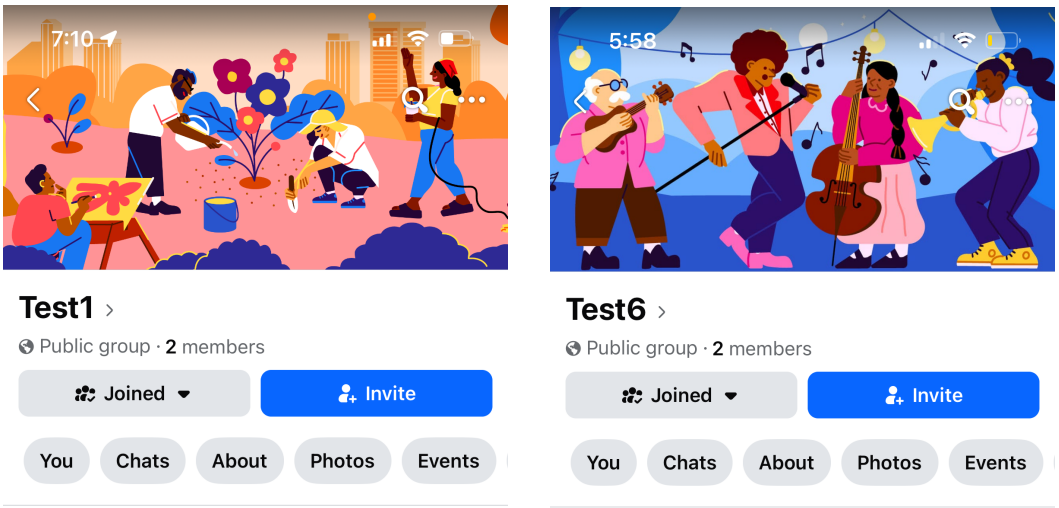


Fig. 2. Representative Facebook Group cover-photo screenshots illustrating the three-dot "Group tools" menu icon in the upper-right corner.

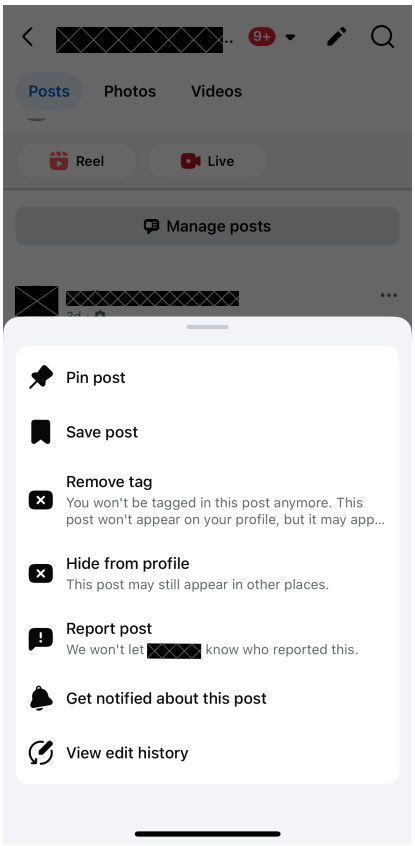
multi-step workflow. Challenges centered on *unexpected focus order*, *confusing interface cues*, and a *lack of progress indicators*—producing learning, compliance and psychological costs based on the instance.[13]

Unexpected Focus Order: VoiceOver users depend on a consistent and predictable focus order when navigating mobile apps. Typically, when a new screen or dialog appears—such as after selecting a menu item—VoiceOver places the initial focus on a prominent element, often the "Back" button or the first actionable control. From this anchor point, users swipe right or left to move linearly through the interface. In our study, participants expected this behavior within moderation flows such as reporting a post or adjusting content settings.

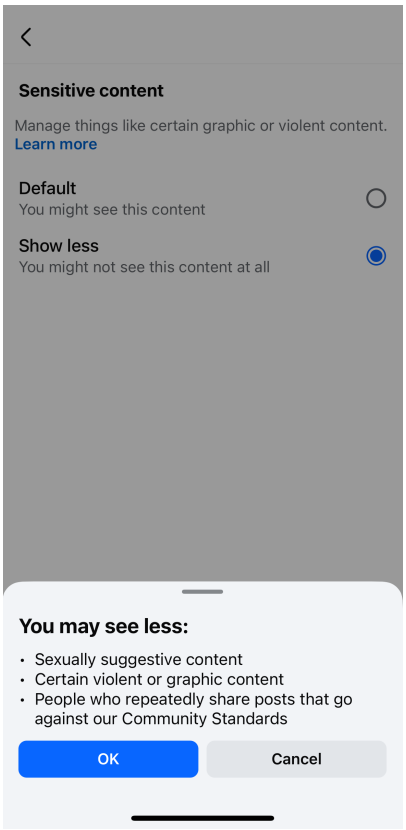
These expectations often broke down in key moderation flows. We observed two types of unexpected focus order: (i) failure to focus on new views and (ii) illogical ordering of existing UI elements. As an example of the former, after choosing to report a post on Facebook, a panel slid up from the bottom (Figure 3), yet VoiceOver failed to focus on it. Instead, VoiceOver focus floated unanchored, announcing only "Dim background" when users touched the screen's upper regions where content is typically expected. The actual controls were at the bottom, but without cues or automatic focus, participants were left disoriented.

Our participants struggled with this focus failure. Some mistook the "Dim background" overlay for a real control and kept tapping it. Others realized it was inert but kept searching for the real control on the screen. A few became completely stuck and restarted the task. **Of nine Facebook participants, eight** VoiceOver users encountered this problem at least once; only Nicole, who used a magnifier, did not. Confusion often led to self-blame before the design flaw became clear. As Samuel reflected after frustration: *"I thought it was my mistake, but it's really the design."* These episodes, borne of both *learning costs* (discovering the tools) and *compliance costs* (navigating through

[13]These costs do not necessarily arise in a fixed sequence or map directly onto particular interface barriers. The same barriers could impose different costs on different participants, whose encounters with these barriers also varied. An illustrative flow diagram in the Appendix (Figure 8) shows how these costs could interact and accumulate during a safety task.



(a) In-feed "Report post" action sheet that the user faces when they click on the "More Options" menu icon.



(b) "Adult content" preference panel that the user faces when trying to change the amount of recommended sensitive content.

Fig. 3. Examples of pop-up panels that failed to auto-focus with VoiceOver, leaving users disoriented during key moderation and safety tasks.

the reporting flow effectively) show how a lapse in focus transfer cascades into disorientation, self-doubt, and repetition—an instance of the platform's *psychological cost.*

Focus and navigation problems extended beyond this initial hurdle. Once participants succeeded in latching onto the pop-up panel, the focus engine still betrayed a sight-first logic, especially on X[14]: it frequently skipped the first actionable element and jumped straight to what a designer might consider the "primary" control. The reporting wizard begins with a prompt—*"Which type of issue are you reporting?"*—followed by a short help link labeled *"Why are we asking this?"*, as shown in Figure 4. When Cristian clicked on that link to learn more, VoiceOver did not land at the explanatory text; instead it leapt past the paragraph and settled on the final *"Got it"* button. Cristian expressed: *"Wait, I asked for the explanation and it's already telling me 'Got it'?...I never heard the explanation!"*

[14]We document the entire workflow for reporting a dummy post using VoiceOver here: https://figshare.com/s/20940fa4b9f7d51c3ad9

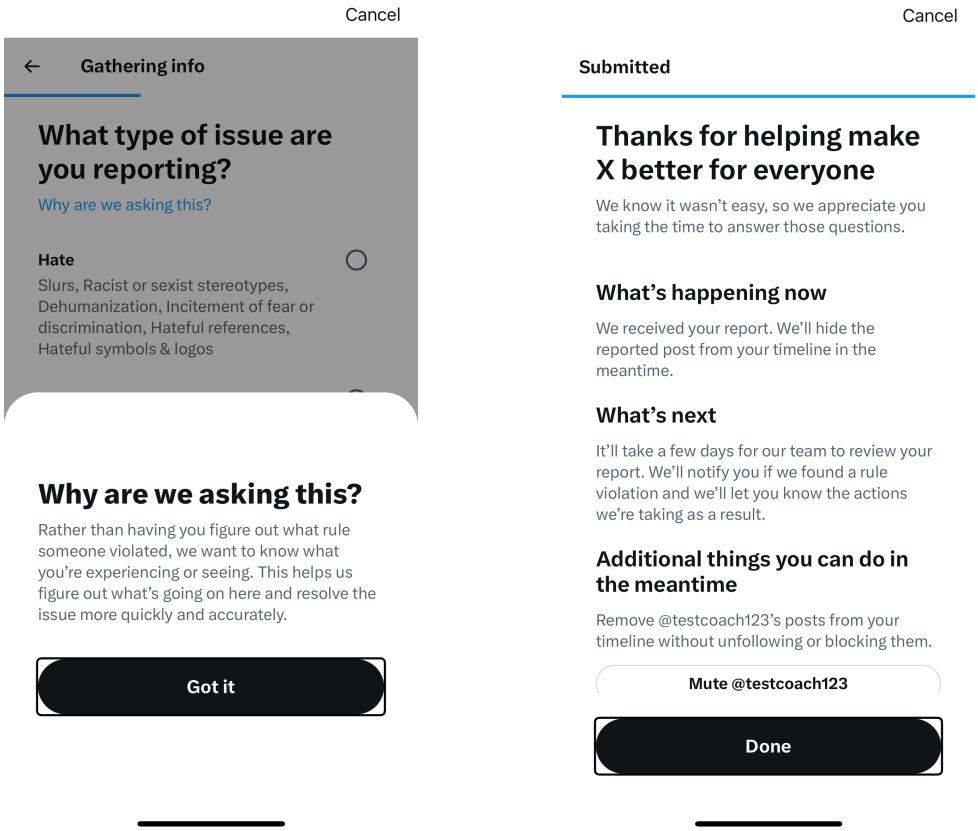


Fig. 4. X's pop-up panels: VoiceOver focusing directly on primary controls in additional-information wizard (left) and in a post-submission confirmation page (right).

A similar focus shortcut appeared at the *end* of X's reporting flow. After Mateo submitted a hate speech report, the confirmation screen offered follow-up actions—mute, block, or limit notifications—stacked above an unmistakable *"Done"* button. VoiceOver, however, landed directly on *Done*. Unaware of the additional safeguards, Mateo simply double-tapped and exited: *"It said 'Done button,' so I pressed it...So you're telling me there's more options further up the page?"* Mateo's experience illustrates what Bigham et al. [4] described as the problem of "not knowing what you don't know": screen reader users cannot perceive elements the interface never announces. In safety workflows, this manifests as missed opportunities for protection—users complete reporting flows unaware that additional safeguards (mute, block, limit notifications) existed but were never surfaced. Through a blend of *learning* and *compliance* costs, users with BLV miss out on critical opportunities to fully engage with the breadth of safety tools available to them.

Focus oddities also amplified the cost of mistakes. Reporting wizards usually require several categorical choices across different interface pages; if a user realized on page 2 that they had picked the wrong category on page 1, there was often no reliable way back. X's workflow, for instance, omits a *"Back"* link after the first step, so changing an earlier answer meant starting the report again, a *compliance cost*. Grant discovered this the hard way and had to abandon the wizard entirely: *"I tapped the wrong reason and...[tried to go back but] found no back button...had to start from scratch."*

Even escaping the reporting wizard, once initiated, was found to be tricky on X. Many participants sought a *"Cancel"* or *"Close"* control in the upper-left corner—mirroring standard iOS convention—only to find the control tucked away at the upper-right on X but announced *last* in the swipe order.

Collectively, these focus and navigation hiccups turn minor deviations from platform conventions into a measurable burden: they create *learning costs* by obscuring where focus should land, increase *compliance costs* by forcing extra swipes, detours, and restarts, and add *psychological costs* through confusion and self-blame, often breaking the user's momentum.

Redundant Cues and Hidden Confirmations: X's reporting wizard presented a different challenge: *compliance costs* built through excessive verbosity coupled with a lack of audible confirmation for selected options. When navigating the category choices while reporting a post, VoiceOver exhibited a redundant pattern: it first read the category heading, then its descriptive sentence, and finally repeated *both* the heading and description together, appending "radio button." The layout of the reporting wizard created three problems:

First, the wizard required users to swipe through long, repetitive blocks of text to reach the "Next" button—a slog Cristian likened to *"wading through every block before you reach 'Next.' "* Because the section headings could not be skipped by VoiceOver without going into the description, participants could not jump by headings and thus had no quick way to skip ahead. The route into the wizard was long as well: *"So it takes 10 swipes down to get to more actions,"* Cristian noted, a *compliance cost* users pay every time they report.

Second, compounding this slowness was a critical lack of auditory feedback. When a participant double-tapped to select an option, the interface accepted the input but offered no audible acknowledgement like "checked." Mateo explained the resulting doubt: *"I tapped while it was still explaining, and it never said 'checked'...so I had no idea it took my answer. You can see it though right?"*. We observe going back to check whether a selection had registered as a compliance cost, and having to learn where the selection was shown as a learning cost.

Finally, the redundant structure of the announcement itself created confusion. After hearing the same content spoken twice, participants sometimes mistook the repetition for a new, distinct category. Cristian echoed this: *"It read the title, the paragraph, then both together. The second time it read it I thought it was a different category...I didn't pay close attention to it until I went back to the description again."* This combination of forced verbosity, missing confirmation, and confusing repetition charged a significant *compliance cost*, and some participants found the repetition overwhelming, a *psychological cost.*

Uncertain Progress: Because neither platform clearly states how many screens lie ahead, first-time reporters often mistook mid-flow pages for the end of the process. Frank, who performed the task on Facebook, admitted, *"If I'm a first-time user...I would just find 'Close' and not realize they're asking me [another question]."* Adriana echoed this sentiment, saying she would prefer cues like *"Step 1 of 3"* upfront so she could gauge whether a post was *"worth the effort of reporting."* X does show a progress bar, but as a percentage; some participants failed to map "33%" to "one-third done." Grant thought the concept was "okay" yet suggested that question or page counts—"1 of 3"—would be clearer. Without such cues, participants grew weary; Adriana confessed on the third page of reporting the subcategory, *"I'm annoyed...I will put my phone down [at this point] and leave it alone."* Her frustration reflects a well-documented usability problem: interfaces that fail to communicate progress through multi-step procedures erode user persistence [66], a *psychological cost* borne of *learning cost* of not knowing how far along one is. The stakes, however, are not merely task

completion but safety itself—when users abandon harassment reports mid-flow due to uncertainty, platforms lose critical signals while users remain exposed to ongoing harm.

Despite the initial fatigue, persevering participants highlighted several positive design elements that tempered their frustration. The workflows themselves, once located, were sometimes praised for their clarity. Vasu called the flow *"smooth" and "comprehensive"* after finding it; the granularity of options conveyed a sense of legitimacy for him. Jessica appreciated that Facebook's wizard *"asked questions to make sure I gathered all the correct information,"* which suggested reports were taken seriously.

Crucially, clear feedback and confirmation were highly valued. Positional cues like "1 of 5" on X's list option items anchored users and reduced guesswork, while summarizing reporting details just before submission and explicit spoken receipts about receiving the report—like Facebook's "Thanks for reporting this post. Next step..awaiting review..."—assured users like Frank that their complaint was logged. Similarly, a simple closing line on X, "Thanks for helping make X better for everyone," provided a welcome sense of closure for Grant. Finally, participants applauded the ease of access to post moderation controls, such as X's integrated options to immediately *"Mute the person...or block them,"* (Figure 5, right) which Maya saw as a considerate touch that restored user control. These positive examples show that clear cues, granular options, and empowering follow-up actions can significantly improve the user experience, lower administrative burdens, and foster a sense that one's effort can yield tangible results.

4.2 Subverting Cognitive Expectations in User Controls

Beyond challenges in discovering and navigating specific workflows, participants encountered numerous design inconsistencies across platforms which undermined their cognitive expectations. Crucially, this section identifies a different set of problems than those outlined above: instead of speaking to issues in the interface itself, this section speaks to the ways that participants' agency was diminished through inconsistency. This inconsistency undermined participants' cognitive expectations of how design elements *should* work—expectations which have been built through regular use of platform interfaces—creating a heavier-than-necessary burden for those with vision loss. Some key issues were *inconsistent labeling*, *lack of preview information*, and *illogical placement and inaccessibility of controls.*

Participants experienced this friction from the moment they attempted to open options menus. A recurring theme was the inconsistent labeling of the overflow menu icon across different contexts on Facebook. While the same "•••" icon is used universally for post, profile, and group options, VoiceOver announced a different label in each case – i.e., "More options" on a post, "See more" on a profile, and "Group tools" on a group page. This lack of uniformity caused confusion and extra mental effort. Laverne captured the mental tax this imposes:

> *"[For profile] It says 'see more.' And the other one [post] was 'more options'... And for another... 'group' something...So with three completely different things [for the same icon], you would have to get comfortable and remember at my age."*

Several participants found this especially frustrating because it imposed more uncertainty for users with BLV compared to sighted users. Vasu defended the context-specific labels at first, but ultimately conceded that a similarly sounding descriptive label would be far better for everyone:

> *"If they call it 'Tools' or 'More options' everywhere, that would have been the best... When you [the interviewer who is sighted] see it you know immediately [that it is an option for menu items]. It should be the same for us."*

Other participants noted that even when they found the icon, the surrounding context did little to signal its purpose as a menu button. As Adriana explained after swiping past the menu icon without recognizing it, *"'See more' doesn't tell me anything. It just says 'See more'… see more of what?"* Adriana's confusion stems from a fundamental interaction design failure: inconsistent labeling of functionally identical elements [67, 70]. For screen reader users, these label variations mean relearning the entry point for every moderation action—reporting, blocking, adjusting privacy. In contexts where rapid response to harassment matters, such friction may determine whether users act at all.

When an action as fundamental as opening an options menu is obscured by inconsistent or opaque labeling, users with BLV must either rely on trial-and-error exploration or abandon the feature altogether. Such friction forces users to learn each label by trial and error, a *learning cost*, and slows task completion, a *compliance cost*. The inconsistency also left participants feeling that the interface was not built for people who navigate without sight, a *psychological cost*.

Once participants finally reached the moderation options, they still had to gamble on what each feature (like flagging, blocking) would actually do. Neither platform previews the consequences: X surfaces an explanation only after the button is pressed, while Facebook reveals the details only once the action is complete. In that vacuum, everyone based their understanding of moderation features, often incorrectly, on their mental models of similar tools they had encountered in other contexts. For example, Laverne (63, who mostly relies on iMessage and phone calls) tried to unblock a user on Facebook by hunting for the person's profile[15]—mirroring how she would unblock a contact on her phone—because, as she reasoned,

> *"I thought it'd be like blocking a number: if I block someone they can't text me...[but] I can still message them. I have no idea how blocking [on Facebook] works."*

Jessica carried a different assumption:

> *"I didn't know you could block someone who isn't your friend... I'd never have learned that if we hadn't done this task."*

Lacking immediate, task-time descriptions, participants built inconsistent mental models of *Block*, *Mute*, and *Report*, discovering the truth only through trial and error—a design gap that left every user guessing about critical safety controls.

Most participants' first instinct when asked to "block the (fictional) hateful user" was to look for the control right on the offending post itself: **eleven of fifteen** swiped through the X or Facebook post, opened the "More options" menu, and were surprised when no *'Block'* command appeared. "*I am seeing if there's [a possibility to] block user from the post...I thought they would have it here*," Patricia complained after swiping through all the options provided on the menu options in the (fictional) hateful post before realizing she had to detour to the author's profile. X technically offers the affordance many wanted—a block button nested in the actions attached to every post—but discoverability was uneven. One participant, Mateo, celebrated it as exactly what Facebook lacks: "*I love that they put the block right there...somebody got you so aggravated...block right there*." Yet others missed the same shortcut because it sits inside a secondary submenu inside the "More actions" menu (see Figure 5). Shreya's first visual sweep through the post options on X produced only silence—"*Oh, there's no option for block from the more [actions] on the post itself*"—and Cristian concluded the feature was "*a little hidden when it could be under more actions [on the post itself]*."

Even more elusive than post-level shortcuts were the platform-wide settings themselves. During tasks requiring participants to *unblock a previously blocked user* or to *adjust sensitive content filters*, **nine of fifteen** participants failed to locate the controls, as they were hidden deep within sprawling

[15]Unblocking a person on Facebook is currently only possible by finding the exact option in the privacy settings.

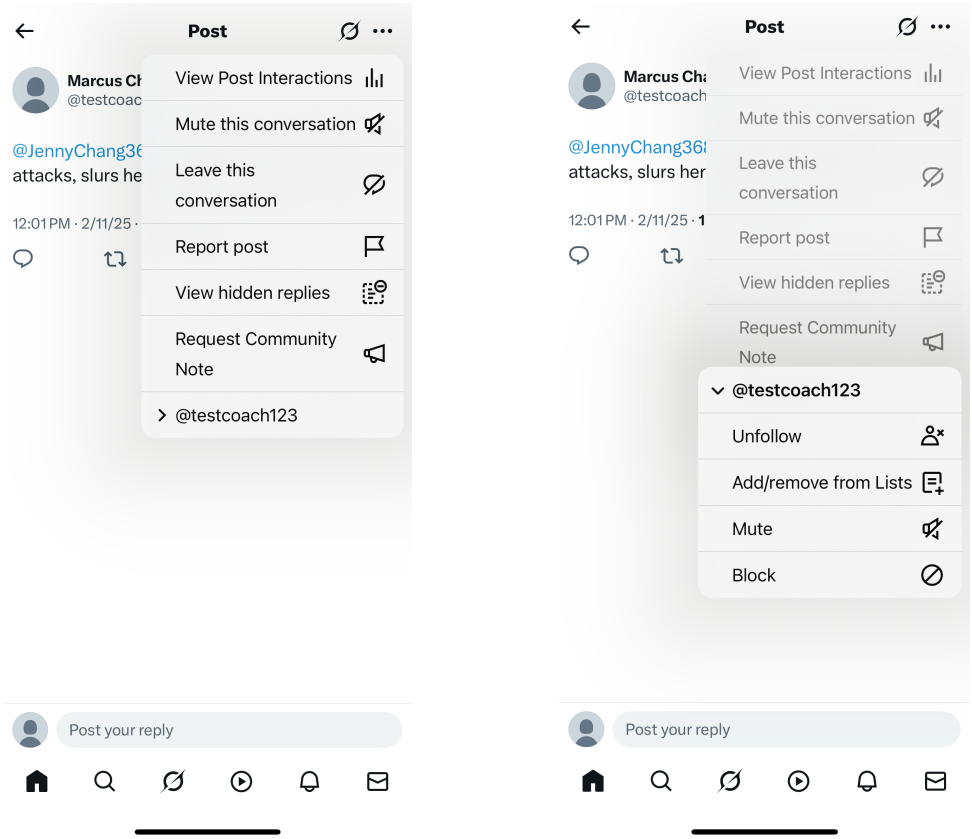


Fig. 5. On X, the "Block" feature is two layers deep: users must first open the post's *More actions* menu (left) and then expand the ambiguous @username submenu (right) before the option appears—hiding a critical safety control along a path that offers little contextual cue for VoiceOver users.

category trees. Jamila quipped that finding them felt like *"hunting for an Easter egg."* The in-app search function, which many turned to as a last resort, offered little help by failing to interpret user intent. For example, on Facebook, participants attempting the unblocking task searched for "unblock" in the privacy settings search bar, which yielded no results. The correct term, "block," was required to find the list of blocked users. Adriana called this literal search function *"stupid,"* as it could not infer her goal. The *learning costs*, in this case, are made paradoxically more burdensome through the in-app search functions meant to support with just such tasks.

A parallel failure played out on X. When tasked with changing their settings to see less sensitive material, participants entered logical search terms like "Adult content," "Sexually explicit," and even "Sensitive content." Yet the search failed to return the relevant toggle, which was filed under the precise label "Display media that may contain *sensitive content*." These rigid and ineffective search mechanisms made platform moderation affordances feel not just hidden, but functionally unavailable to our participants. This search failure exemplifies what Furnas et al. [26] termed the "vocabulary problem": users generate intent-correct search terms ("unblock," "adult content") that systems fail to match against their actual labels ("block list," "sensitive content settings"). The consequence in safety contexts is stark—controls that exist become controls that cannot be

found, effectively removing users' ability to manage their protection despite the platform nominally providing these tools.

Taken together, these uncertainties translate into a concrete burden: they create *learning costs*—users must infer meanings for identically-shaped icons across contexts, act without task-time previews, and wrestle with literal, intent-blind search—and they raise *compliance costs* by forcing detours (e.g., from posts to profiles), multi-level submenu traversal, and repeated attempts (e.g., 11/15 first looked for *Block* on the post; 9/15 could not locate the relevant settings). This led to guessing, backtracking, and often abandonment of the very safety tools meant to empower users.

4.3 Adaptation, Emotional Labor, and Perceived Agency

We next turn from pragmatic and functional barriers to users' elucidation of their problems. We document the adaptations participants employed, the extra interaction work these entailed, and the emotional consequences that followed. These consequences span all three costs. Workarounds required participants to learn gestures and outside tools (learning costs), and when they worked, they reduced the number of steps a task took. The extra interaction work raised compliance costs, and the emotional labor and loss of agency that followed were psychological costs.

4.3.1. User Workarounds

Participants described inventive workarounds to navigate content moderation interfaces when standard methods fell short. When users on X swiped past the menu option that reveals more controls (including reporting), they resorted to touch exploration—sliding a finger across the screen surface to identify and interact with elements by vocal feedback.

Cristian—a highly proficient VoiceOver user and a trainer for using assistive technology—bypassed the elusive menu icon on X by placing his finger on a post and performing a double-tap-and-hold gesture (Figure 6, left) (equivalent to a long press) to instantly reveal a context menu with options including "Report Post," a workaround he learned by himself to avoid the lengthy sequential swipes needed otherwise. Because such non-obvious shortcuts are easily missed by novices who rely solely on flicking through content, Cristian now teaches this technique to other blind users, framing it as a vital strategy for uncovering hidden actions within X's interface.

Low-vision participants demonstrated their own arsenal of adaptation strategies. Some preferred magnifying content visually over using screen readers. For instance, Shreya (who has extremely high myopia) relied on iOS's Display Zoom and large text settings to enlarge posts, only toggling VoiceOver on for certain tasks. She expressed a personal attachment to "seeing" content, trying to remain a visual user despite her deteriorating sight. Nicole likewise used a physical 11x magnifying glass with her phone, holding it over the screen to read fine print or interface labels. This workaround helped her avoid the layout disruptions caused by oversized text settings. *"I have to use my magnifier all the time on my phone...it's just habit,"* Nicole admitted, acknowledging that increasing the on-screen font to a comfortable size often made content run off the screen. For her, a handheld magnifier provided a more predictable view of the screen layout that would otherwise be cut off or require panning.

When social media apps proved too confusing even with previously discussed workarounds, participants turned to virtual assistance tools like 'Be My Eyes'[16] (which connects to a volunteer or AI who can describe the screen) or Siri. Cristian mentioned that if he ever truly got stuck, he would take a screenshot of the mobile app layout and ask the 'Be my AI' feature of the 'Be my Eyes' app to describe the screen for him. He also described a personal technique of using a second finger as a physical "marker" on the screen to help track his current position while navigating

[16] https://www.bemyeyes.com

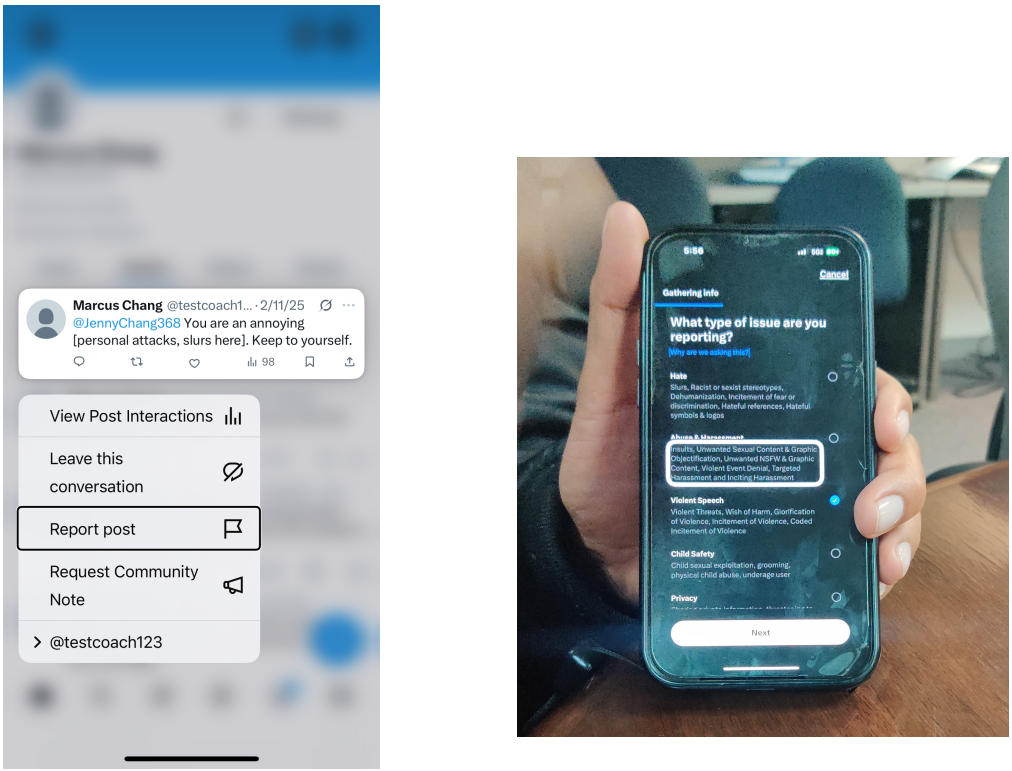


Fig. 6. Workarounds adopted by Cristian while using X: Double-tap and hold brings options available haptic actions (left) and using his middle finger as a marker for the current action line (right).

or checking other information (like notifications or time) (Figure 6, right). Cristian suggested this highlighted the need for built-in features, like a VoiceOver "hot spot" option, to temporarily bookmark a location on the screen and easily return to it after interruptions, reducing the mental labor of re-orienting.

Through techniques like touch exploration, memorizing screen layouts, utilizing zoom and magnification, or seeking human assistance, BLV users exhibited resilience and ingenuity. These adaptations empowered participants to complete reporting and blocking tasks. However, employing these workarounds often required extra effort, expertise, and patience—resources that not all users have in equal measure. This personal labor compensated for design shortcomings, allowing our participants to exert agency over content moderation actions *in spite of* the platforms, rather than because of them.

4.3.2. Interaction Overload and Participant Fatigue

Even with clever strategies, the labor involved in content moderation tasks was strikingly high. Participants frequently encountered workflows that were labyrinthine, requiring dozens of gestures and trial-and-error attempts to accomplish what seemed like a simple goal. For instance, changing a privacy setting or unblocking a previously blocked user could demand navigating through multiple layers of menus, each involving sequential swipes, double-taps, and back-tracking through screens. We observed that without using the search function to navigate directly to a setting, users might

have to perform in the order of *thirty* or more discrete actions to reach a deeply nested option, assuming that they don't take detours, which would further exacerbate tasks.

Our participants voiced the *psychological costs* of these drawn-out procedures. Several described feeling as if they were *"walking in circles"* through menu trees and when they found it, they *"didn't even have an idea how they got there"*. One user quipped that she had to attempt the task three separate times initially before learning the correct path to her blocked list. During a Facebook task to unblock a user, Samuel became so frustrated by the maze of options that he eventually asked Siri for help, treating the problem as one big question to be answered externally. He confessed that without an external shortcut, he felt lost in Facebook's interface: *"I don't have knowledge of the whole interface,"* he said, after combing through Settings fruitlessly. The number of sub-menu categories and lack of clear pointers wore him down. Describing the convoluted menu system, Samuel joked— *"...there are so many options, and more options inside them. They should [instead] call it 'Maze-book' or something."*

4.3.3. Emotional labor and Perceived Lack of Agency

Beyond the practical challenges of interacting with moderation tools described in the previous sections, participants' experiences with reporting and moderation carried an emotional labor. Engaging with these tools – often in the context of hateful or abusive content – sometimes provoked anxiety, distress, and a diminished sense of agency. Several users expressed trepidation about even initiating a report. Laverne, for example, was so wary of "doing something wrong" on Facebook that she hesitated to try any moderation settings at all on her own profile. She had heard rumors about users being penalized for acting in suspicious ways, which made her afraid to experiment or make a mistake. As she put it:

> *"I'm afraid to just touch anything..I don't know what I'm doing. I've heard people say you can get put in Facebook prison."*

Laverne shared that she was on Facebook only to read posts from her granddaughter and was apprehensive about losing access to the platform.

Shreya and Jamila, both of whom had confronted harassment online, described feeling disheartened and even mocked by the outcome of reporting incidents. They had dutifully reported abusive posts directed at them, only to see no visible action taken against the perpetrators. Jamila recounted reporting several hate-filled comments on her post in a local community group, but, *"usually nothing comes of it,"* she said–the toxic posts remained up, and moderators or Facebook staff often responded with silence or a generic rationale about free speech. These experiences echo prior findings on post-reporting grievances of sighted users [106], but they emphasize their particular salience for this especially vulnerable population.

Moments like this imbued a sense of futility. Participants felt that they were doing their part by courageously flagging abuse, yet the platforms did not reciprocate with adequate support or resolution. Shreya shared that seeing rampant harassment go unchecked on X made her withdraw from interacting altogether–*"why put myself at risk of further harassment if it [X] would not back me up?,"* she reasoned.

This emotional disillusion eroded the trust that reporting mechanisms would actually help; instead, using them sometimes felt like reopening wounds for no benefit according to some participants. In addition to the *psychological burdens* of fear and disappointment, many participants voiced frustration at the false sense of agency created by these tools. Social media platforms outwardly empower users to classify and curb harmful content, but in practice our participants often felt like the burden of moderation was shifted onto them without real power. Participants bristled at being forced to categorize the offense themselves. After navigating a reporting form on Facebook, Vasu

questioned why the onus was on him to decide if something was "harassment" or "hate speech," asking rhetorically, *"Why do I have to do it? Isn't it their job?"* Patricia felt that forcing a user to select the precise category of abuse was asking them to *"endure [the pain]...for your [platform's] convenience."* In her eyes, the reporting process was designed for the platform's convenience rather than the user's well being. These reactions mirror prior work showing that flagging/reporting systems conscript end-users and volunteer moderators into governance roles while platforms provide limited follow-through or support—leaving people to shoulder moderation labor with little feedback or remedy [15, 44, 60].

Encounters with illusory controls—such as X's *Report community* link that merely redirects[17] to a help article—further eroded participants' faith that moderation tools would actually address their concerns. Several participants suspected their submissions served mainly to refine platform's content curation algorithms rather than to secure personal relief. In sum, we documented our participants' *psychological cost* of this ambiguous labor: anxiety over potential backlash, discouragement when no feedback arrived, and anger at having to police abuse without meaningful support. While prior work documents that reporting systems offload labor onto users [15] and often yield no action [44, 106], our findings reveal an accessibility-specific double burden. Users with BLV like Shreya and Jamila—who already face disproportionate harassment [40, 54]—must first navigate unlabeled buttons and opaque workflows to file reports, then experience the same futility sighted users face. This compounding of accessibility work and ineffective outcomes helps explain participant withdrawal from protective actions entirely.

5 Discussion

We discuss our results through two intertwined manifestations of administrative burden in end-user safety tasks for participants: *uncertainty* and *labor*. By uncertainty we mean lack of clarity in situations where participants have to figure out whether safety controls exist, where they are, and what they will do, which disrupts mental models and progress. By labor we mean the effort users expend to make sense of a situation and reduce uncertainty so they can proceed—repeating gestures, taking detours or restarts, and devising workarounds or seeking help. We foreground these aspects because they recur across our data and let us trace how administrative burden's *learning, compliance and psychological* costs, detailed in section 4, interleave in practice. The subsections that follow explicate where these costs emerge, intertwine, and cascade across steps, and we provide suggestions to mitigate these costs.

5.1 Uncertainty: Learnability, Opacity, and Mental-Model Drift

We begin with *uncertainty* because, in our data, not knowing what a safety control does, where it lives, or how the control functions persistently incurred *learning costs* at the outset of safety work. These learning costs then spilled into *compliance* and *psychological* costs as people probed, backtracked, and second-guessed their actions. In our analysis, uncertainty surfaced at four recurrent points that we detail below: (1) whether a control existed at all; (2) inconsistent names applied to the same icon; (3) missing cues about how the tool will function before commitment; and (4) vague or misleading progress cues; Together these form end-to-end transparency gaps—a mix of *semantic opacity*—where inconsistent labels and intent-blind search undercut discoverability, and *procedural opacity*—where users get little insight into what a control will do and only vague signals of progress, completion, or outcome. These gaps disrupt users' mental models and transform safety work to precarious guesswork.

[17] A demonstration of this redirection is provided here: https://figshare.com/s/4be546a24b0610accde7

(1) Existence uncertainty. Participants were often unsure whether certain safety functions were available at all, especially for low-salience privacy and audience controls. Plain-language searches in the 'Settings' menu commonly failed due to keyword-only matching and narrow indexing, so intent-correct terms (e.g., "unblock") did not surface the relevant capability ("block list"). This pattern reflects the *vocabulary problem*—people naturally use diverse, intent-correct words that do not match system labels—forcing incremental probing to learn whether a capability exists [26, 86].

These conditions elevate *learning* costs (time spent hypothesizing and trying variants), spill into *compliance* costs (detours through help centers or off-app searches), and add *psychological* costs (doubt and early abandonment). Accordingly, the remedy is *intent-aware search*: normalize synonyms and near-misses, and surface all semantically relevant controls even when queries do not match system labels [26, 36].

(2) Inconsistent labeling. Identical icons or entry points carried different labels across views, e.g., "More options," "See more," "Group tools." For screen-reader users, these renamings break learned mappings, shift effort from recognition to recall [10], and increase mis-selections and extra gestures. In practice, participants ended up re-learning labels (*learning cost*), and recovered from avoidable mis-selections—this necessitated extra taps and backtracking that introduced hesitation and slowed the reporting task (*compliance cost*). To address such problems, we recommend that platforms align their web interfaces with the Web Accessibility Initiative's principle of "Consistent Identification" [97]. Using the same label for the same function reduces re-learning overhead and keeps non-visual navigation predictable, echoing broader usability heuristics that emphasize recognition over recall [10].

(3) Missing functional cues. On Facebook, the 'Block' control offered no foresight into what activation would entail—and, once invoked, it provided no auditory confirmation that the action was completed; instead, the blocked person's profile was thereafter hidden from the participant's interface, with no indication that this outcome was expected. On X, the pattern was similar: neither 'Block' nor 'Mute' provided upfront descriptions of their consequences, and explanations appeared only after completing the task, on the confirmation screen. These missing cues widen Norman's *"Gulf of Execution"*—uncertainty about whether the available action matches the user's intent—and *"Gulf of Evaluation"*—insufficient feedback to interpret the new state [67]. A straightforward fix is to provide concise, screen-reader-exposed previews before commitment (e.g., "Blocking x hides their replies and prevents mentions...") and immediate, focused auditory confirmations after commitment (e.g., "You have successfully blocked x...the profile will no longer be visible to you on the platform"), so intent, action, and consequences align in a continuous, understandable flow.

(4) Progress ambiguity. Participants repeatedly expressed disorientation caused by unclear workflow endpoints and ambiguous mid-task signals. For instance, the lack of a progress indicator in the reporting wizard on Facebook led to premature exits among some participants. According to Tversky and Kahneman's framework of decision-making under uncertainty [89], ambiguity amplifies perceived effort, discouraging continued engagement. Thus, ambiguous or incomplete feedback loops in moderation interfaces directly contribute to higher attrition among users with vision loss, intensifying their administrative burden through unnecessary cognitive labor and interaction fatigue. Clear step structures ("Step 2 of 4"), named stages, and immediate confirmations implement the classic visibility-of-system-status heuristic championed by the UX expert Jakob Nielsen [35, 66].

In sum, uncertainty around labels, previews, progress, and outcomes did not merely slow down task initiation; it eroded momentum toward completion, making later steps costlier and more fragile. When interfaces did not resolve that uncertainty, participants compensated with *work*—detours, repetitions, and restarts—to keep going. The next subsection examines this labor and its characteristic accumulation patterns.

5.2 Labor: Inequitable Interaction Work

Our data show that participants incurred substantial *compliance costs*—the time and effort spent on tedious chores required to execute safety tasks—accumulating through two recurrent mechanisms. First, *carry-over costs*: when the first step was unclear or hard to reach, extra work cascaded into later steps. Second, *restart costs*: when obvious cues to continue or exit ("Back" or "Cancel" buttons) weren't readily accessible, or when VoiceOver did not focus on any element or focused on the wrong control, people often had to quit their flow and start over. While these mechanisms co-occurred with learning costs (e.g., locating or interpreting controls), our emphasis here is on how interaction work itself accumulated and exacerbated psychological spillovers (hesitation, self-doubt, fatigue).

Carry-over costs. When the entry point is hard to find or ambiguous, the extra work does not end with discovery; it propagates. Participants who could not reach a reliable entry (e.g., menu items skipped in linear focus order, icons requiring magnification and panning for people with low-vision) accumulated more exploratory gestures upfront and faced longer downstream sequences to reach the same endpoint. The result is a shift from a short, linear path to a branching search-and-verify routine that invites more compliance cost in the form of attention and time.

Even after discovering the correct starting menu icon, safety controls are buried deep within menu trees, and progress requires long chains of micro-actions that redirect attention from judgment to wayfinding. Even on the correct path, reaching a single control can take *dozens* of gestures without search. Unskippable panels and verbose wizards exacerbate the interaction effort required to complete the task. This amount of work turns routine execution into accumulating *compliance costs* while draining users' *psychological reserves*, resulting in fatigue, frustration and attrition.

Restart costs appeared when participants lost their place or the system offered no path backward. On X, the category selection wizard offers no "Back" control, so changing an earlier answer meant restarting the task. Loss of automatic VoiceOver focus, unintuitive placement of 'Back' and 'Cancel' buttons in the navigation sequence, and ambiguous progress indicators repeatedly forced participants into abandoning their progress and restarting the task. Restarts are not a single penalty: they erase context, invalidate partial progress, and multiply opportunities for new errors on re-entry. In time-sensitive scenarios (e.g., escalating harassment), these resets translate into missed opportunities for timely intervention, and erode users' willingness to continue reporting, increasing drop-offs.

We read the extra work described above as labor extracted from users to keep safety flows moving. Terranova's notion of *free labor* [87] clarifies how platforms benefit from users' enforcement inputs (reports, blocks, classifications) through unpaid work. The burden of this labor is exacerbated for those navigating these controls through inaccessible channels, thus creating an inequitable distribution of the everday burden of staying safe online [87]. Read together—and situated within the administrative burden of safety work—these frames clarify who pays for the surplus work, why it is inequitable, and how this free labor is unaligned with design-justice commitments to distribute benefits and burdens fairly [14].

Participants also questioned the fairness [59, 91] of shouldering *governance work*: not only executing a flow but translating lived harms into platform taxonomies, interpreting policy language, supplying evidence (e.g., screenshots, timestamps), and then tracking outcomes across scattered surfaces. This aligns with critiques that flagging shifts governance work onto users, who must shoulder the labor of complaint [15]. In practice, this translates into point-of-use "paperwork": users must select categories that may not match lived harm, justify choices without previews of consequences, and then monitor opaque channels for outcomes—administrative steps that grow when platform-side redress infrastructure is weak. While prior work notes that much of the burden

of staying safe online still falls on users themselves [48, 100], our findings show that this burden significantly exacerbates for users with low vision.

5.3 Implications

To mitigate this inequitable labor, we offer implications for different stakeholders.

1. **Legislators**. Accessible online safety tools require enforceable accessibility obligations. Our analysis found that Facebook's menu icon fell below the WCAG 3:1 contrast benchmark on five of six default group cover images, as did X's post-menu icon in its default configuration. Several jurisdictions, including the EU, require digital services to meet accessibility standards based on WCAG. In the EU, this benchmark is incorporated into EN 301 549 [19]. Whether these failures also constitute legal violations depends on the applicable obligations: the European Accessibility Act covers specified services, including e-commerce and electronic communications [21]. Our findings illustrate the stakes of the European Disability Forum's warning about the EU's Digital Services Act (DSA), which governs online platforms, that *"the mechanisms to flag hate speech or illegal goods will only be accessible depending on the good will of the digital services providers"* [20]. The final DSA addresses accessibility for persons with disabilities through codes of conduct (Article 47), which it frames as voluntary (Article 45(1)) [22]. We therefore recommend that legislators explicitly require accessible reporting mechanisms, with enforceable criteria covering the visibility and screen-reader accessibility of entry points, navigation through reporting forms, and submission feedback. Our study contributes to a strong line of research on the disproportionate safety risks that users with disabilities face online [40, 41, 54, 56], and serves as further evidence that they do not have equitable access to the tools meant to empower them.

Given that this study is situated in the U.S. context, the Americans with Disabilities Act (ADA) is also worth invoking when considering legal accountability. Although there have not been uniform technical accessibility standards forwarded by the U.S. Department of Justice, websites and mobile applications of "public accommodations" (i.e., businesses that supply goods and services to the public) must be accessible to users with disabilities under Title III of the ADA. While the applicability of this rule for entities without brick-and-mortar storefronts is contested, practitioners writing in the American Bar Association's *Business Law Today* advise that all sites and apps comply to avoid litigation [79]. We would argue that there is a very real way in which the lapses in the accessibility of safety tools outlined in this study are, in effect, undercutting the digital accessibility rights of users with BLV and ultimately making them less safe than their sighted counterparts online. For that reason, frameworks such as the ADA could potentially be invoked to force the hand of platforms in attending to these concerns.

2. **Interface designers**. Platform designers must shift from a sighted-default paradigm toward proactive and ability-centered design [101]. *Proactively surfacing moderation controls with adequate context* directly on posts can substantially reduce navigational complexity. Clear and intuitive navigation structures, correct focus placements for the screen reader, and readily accessible 'Back' and 'Cancel' options are essential for streamlining interactions and preventing unnecessary detours. For low-vision users, controls must also remain visually distinguishable across backgrounds. One participant, Nicole, suggested placing a contrasting box behind the menu icon to keep it visible over group cover photos. Building on this suggestion, platforms could provide a solid contrasting background or adapt icon colors to maintain sufficient contrast as the background changes.

Recent research [24, 92] suggests how LLMs are increasing user accessibility across different social media platforms. Our line of work can be mutually informative with this research: while LLMs help resolve accessibility breakdowns with higher efficiency, our approach identifies scenarios and contexts of the administrative burden therein. Although LLMs present a great opportunity to bridge the social-technical gaps [2] experienced by users with BLV, designers should take care to integrate

LLM capabilities in digital platforms in ways that do not further exacerbate the administrative burdens of using them.

3. **Platform policymakers**. Platforms have a clear responsibility to rectify this issue, particularly as they are electing to increasingly push safety work onto users. They must fundamentally change their design processes to embrace participatory design methods that meaningfully involve users with vision-related disabilities at every stage of development. This participation would ensure that critical accessibility issues are identified and resolved before becoming entrenched in the infrastructure. In line with the disability rights principle *"Nothing About Us Without Us"* [14], this means compensating contributors with disability for their expertise and incorporating them as co-designers and decision-makers. Doing so moves responsibility for accessibility upstream, reduces the accumulating administrative burden of safety work, and establishes clearer lines of accountability. These structural commitments—surfacing controls with context, streamlining navigation, and involving disabled users in design—address administrative burden at its source by reducing *learning costs* (clearer discovery), *compliance costs* (fewer steps), and *psychological costs* (greater confidence and agency).

4. **Users**. Finally, users share stakeholder interest in this problem, as they have the power to create opportunities for inclusive community knowledge-building. Users with vision loss are a particularly powerful support in this context, as they can provide tutorials and other forms of collaborative information-sharing for those with similar disabilities. Research has shown particularly promising results on how blind people can collaborate with a diverse group of people (blind and/or sighted people) to address accessibility barriers [56]. Sighted users, too, hold a critical place in this opportunity for community-building. They can, at baseline, be vocal advocates for inclusive design principles in social media spaces, but they can also work to promote and cultivate a more supportive online environment for those who experience moderation breakdowns. Such peer support should complement, not replace, platform responsibility; otherwise it adds to the free labor we describe in Section 5.2.

5.4 Limitations

While this study provides in-depth qualitative insights into the experiences of blind and low-vision users with content moderation tools on two major social media platforms, we acknowledge certain limitations inherent in our research design and scope. Our investigation, primarily focused on specific user-enacted moderation tasks within Facebook and X, does not represent an exhaustive audit of all accessibility and usability issues across every moderation feature or all social media platforms. The tasks themselves, designed to cover a range of moderation functionalities such as reporting, blocking, and adjusting privacy settings, prompted engagement with tools that participants might not frequently or ever use in their daily routines, potentially not fully capturing organic interaction patterns or motivations. Similarly, our text-based placeholders supported examining reporting workflows through task completion but did not capture the perceptual, contextual, and emotional demands of interpreting in situ harmful multimedia content, which may also influence reporting decisions and navigation. However, we note that the ready usability of such tools becomes critical amid safety concerns, and therefore, tool designs must especially accommodate first-time users.

The findings reflect a snapshot in time, capturing the state of these platforms, their respective app versions, operating system configurations, and device models encountered during the period of data collection. Given the highly dynamic nature of these digital environments, where interfaces are frequently updated and A/B testing is common, some observed issues may have since been altered, and indeed, we noted instances where specific challenges were not uniformly replicable across researchers' own devices or every participant's encounter. Our primary methodological aim was to highlight the patterns our participants faced and identify systemic barriers during these

tasks, recognizing that individual encounters with these issues varied due to differences in personal assistive technology proficiency, adaptive strategies, and the aforementioned platform dynamism. However, our documentation of such systemic barriers, despite the wide popularity and age of both platforms suggests a long-standing, industry-wide apathy toward accessibility concerns.

6 Conclusion

This study demonstrates how the design of moderation tools on major social media platforms places an unfair burden of safety labor on users with vision impairments. In order to achieve more equitable access to content moderation work flows, and to better protect marginalized groups online, an array of stakeholders should take these findings seriously: legislators, interface designers, platform policymakers, and users among them.

In all, rectifying the problem of inaccessible safety work goes beyond technical fixes. It requires a genuine commitment to enacting procedural fairness in design across stakeholder groups to make online safety achievable for everyone—not just sighted users.

Acknowledgments

We are extremely grateful to our participants for their time, insights, and help recruiting others. We thank the New Jersey Commission for the Blind and Visually Impaired (CBVI) for sharing our call through their mailing list, and Professor Britt Paris for feedback and helpful discussions. We also thank the East Orange Public Library, the Maurice M. Pine Free Public Library in Fair Lawn, the Archibald S. Alexander Library at Rutgers–New Brunswick, the Carteret Public Library, and CBVI's Administrative Office in Newark for permitting us to use quiet spaces for interviews. This work was supported by the National Science Foundation under Grant No. 2329394. Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the NSF or CBVI.

References

[1] World Wide Web Consortium (W3C) 2025. *Web Content Accessibility Guidelines (WCAG) 2.1.* World Wide Web Consortium (W3C). https://www.w3.org/TR/WCAG21/

[2] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[3] Cynthia L. Bennett, Jane E, Martez E. Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173650

[4] Jeffrey P. Bigham, Irene Lin, and Saiph Savage. 2017. The Effects of "Not Knowing What You Don't Know" on Web Accessibility for Blind Web Users. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Baltimore Maryland USA, 101–109. https://doi.org/10.1145/3132525.3132533

[5] Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. 2013. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1225–1236. https://doi.org/10.1145/2441776.2441915

[6] Stacy M. Branham and Shaun K. Kane. 2015. The Invisible Work of Accessibility: How Blind Employees Manage Accessibility in Mixed-Ability Workplaces. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 163–171. https://doi.org/10.1145/2700648.2809864

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

[8] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[9] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (2021), 328–352. https://doi.org/10.1080/14780887.2020.1769238

[10] Raluca Budiu. 2024. *Memory Recognition and Recall in User Interfaces.* https://www.nngroup.com/articles/recognition-and-recall/ Nielsen Norman Group. Accessed: 2025-08-22.

[11] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[12] James I. Charlton. 1998. *Nothing About Us Without Us.* University of California Press, Berkeley. http://www.jstor.org/stable/10.1525/j.ctt1pnqn9

[13] Bobby Chesney and Danielle Keats Citron. 2019. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review* 107, 6 (2019), 1753–1819. https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security

[14] Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need.* MIT Press. https://doi.org/10.7551/mitpress/12255.001.0001 Open access edition.

[15] Kate Crawford and Tarleton Gillespie. 2016. What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society* 18, 3 (2016), 410–428. https://doi.org/10.1177/1461444814543163

[16] Maitraye Das, Darren Gergle, and Anne Marie Piper. 2019. "It Doesn't Win You Friends": Understanding Accessibility in Collaborative Writing for People with Vision Impairments. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 191 (Nov. 2019), 26 pages. https://doi.org/10.1145/3359293

[17] Tom Divon and Jonathan Corpus Ong. 2025. Tech Bro Power Play: Zuckerberg vs. Global Tech Justice. *Tech Policy Press* (9 Jan. 2025). https://www.techpolicy.press/tech-bro-power-play-zuckerberg-vs-global-tech-justice

[18] Azza El-Masri, Martin J. Riedl, Inga K. Trauthig, and Samuel C. Woolley. 2024. Lodging complaints against platform power: how Lebanese journalists and activists experience reporting mechanisms, platform failures, and techno-alienation. *Information, Communication & Society* 27, 12 (2024), 2242–2259. https://doi.org/10.1080/1369118X.2024.2406803 arXiv:https://doi.org/10.1080/1369118X.2024.2406803

[19] European Commission. 2025. Web Accessibility Directive—Standards and harmonisation. https://digital-strategy.ec.europa.eu/en/policies/web-accessibility-directive-standards-and-harmonisation. Last updated: 2025-05-05. Accessed: 2026-02-10.

[20] European Disability Forum. 2022. Digital Services Act fails to ensure accessibility for persons with disabilities. EDF calls the EU to rectify this political agreement. https://www.edf-feph.org/digital-services-act-fails-to-ensure-accessibility-for-persons-with-disabilities-edf-calls-the-eu-to-rectify-this-political-agreement/

[21] European Union. 2019. Directive (EU) 2019/882 on the accessibility requirements for products and services. https://eur-lex.europa.eu/eli/dir/2019/882/oj/eng. Accessed: 2026-02-10.

[22] European Union. 2022. Regulation (EU) 2022/2065 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng. Official Journal of the European Union, L 277, 27 October 2022, p. 1. Accessed: 2026-09-28.

[23] Facebook. 2015. Introducing the Accessibility Toolkit. Facebook Engineering. https://engineering.fb.com/2015/03/31/developer-tools/introducing-the-accessibility-toolkit/ Published March 31, 2015. Accessed September 15, 2026.

[24] N. Fathallah, D. Hernández, and S. Staab. 2025. AccessGuru: Leveraging LLMs to Detect and Correct Web Accessibility Violations in HTML Code. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–22.

[25] Federal Trade Commission. 2025. *Consumer Sentinel Network Data Book 2024.* Technical Report. https://www.ftc.gov/system/files/ftc_gov/pdf/csn-annual-data-book-2024.pdf Reports over $12B in consumer fraud losses in 2024.

[26] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. 1987. The vocabulary problem in human-system communication. *Commun. ACM* 30, 11 (Nov. 1987), 964–971. https://doi.org/10.1145/32206.32212

[27] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.

[28] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media.* Yale University Press, New Haven, CT.

[29] Katerina Gkatzola and Konstantinos Papadopoulos. 2024. Social media actually used by people with visual impairment: A scoping review. *British Journal of Visual Impairment* 42, 3 (2024), 832–848. https://doi.org/10.1177/02646196231189393 arXiv:https://doi.org/10.1177/02646196231189393

[30] Cole Gleason, Patrick Carrington, Lydia B. Chilton, Benjamin M. Gorman, Hernisa Kacorri, Andrés Monroy-Hernández, Meredith Ringel Morris, Garreth W. Tigwell, and Shaomei Wu. 2019. Addressing the Accessibility of Social Media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. ACM, Austin TX USA, 474–479. https://doi.org/10.1145/3311957.3359439

[31] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019. Making Memes Accessible. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 367–376. https:

//doi.org/10.1145/3308561.3353792

[32] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376728

[33] James Grimmelmann. 2015. The Virtues of Moderation. 17, 1 (2015), 42–109. https://openyls.law.yale.edu/mapping/yjolt/vol17/iss1/2/

[34] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 466 (Oct. 2021), 35 pages. https://doi.org/10.1145/3479610

[35] Aurora Harley. 2018. Visibility of System Status (Usability Heuristic No. 1). *Nielsen Norman Group* (3 June 2018). https://www.nngroup.com/articles/visibility-system-status/

[36] Simon Harper and Yeliz Yesilada. 2008. *Web accessibility: a foundation for research.* Springer Science & Business Media.

[37] Ian Hawkins, Jessica Roden, Miriam Attal, and Haleemah Aqel. 2023. Race and gender intertwined: why intersecting identities matter for perceptions of incivility and content moderation on social media. *Journal of Communication* 73, 6 (12 2023), 539–551. https://doi.org/10.1093/joc/jqad023 arXiv:https://academic.oup.com/joc/article-pdf/73/6/539/54463464/jqad023.pdf

[38] Natali Helberger, Jo Pierson, and Thomas Poell. 2018. Governing online platforms: From contested to cooperative responsibility. *The information society* 34, 1 (2018), 1–14.

[39] Nicola Henry and Anastasia Powell. 2018. Technology-Facilitated Sexual Violence: A Literature Review of Empirical Research. *Trauma, Violence, & Abuse* 19, 2 (2018), 195–208. https://doi.org/10.1177/1524838016650189

[40] 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 2024 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. https://doi.org/10.1145/3613904.3641949

[41] 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 *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 173, 18 pages. https://doi.org/10.1145/3706598.3713997

[42] 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. https://doi.org/10.1145/3517428.3544801

[43] Maximiliano Jeanneret Medina, Yong-Joon Thoo, Cédric Baudet, Jon E. Froehlich, Nicolas Ruffieux, and Denis Lalanne. 2025. Understanding Research Themes and Interactions at Scale within Blind and Low-vision Research in ACM and IEEE. *ACM Trans. Access. Comput.* 18, 2, Article 7 (April 2025), 46 pages. https://doi.org/10.1145/3726531

[44] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 192 (2019), 192:1–192:33 pages. https://doi.org/10.1145/3359294

[45] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages. https://doi.org/10.1145/3338243

[46] 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. https://doi.org/10.1145/3491102.3517505

[47] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction* 25, 2 (2018), 12:1–12:33. https://doi.org/10.1145/3185593

[48] Shagun Jhaver and Amy X. Zhang. 2025. 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* 27, 5 (2025), 2930–2950. https://doi.org/10.1177/14614448231217993 Original work published 2025.

[49] Jeongwon Jo, He Zhang, Jie Cai, and Nitesh Goyal. 2025. AI Trust Reshaping Administrative Burdens: Understanding Trust-Burden Dynamics in LLM-Assisted Benefits Systems. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3715275.3732077

[50] J. Bern Jordan, Victoria Van Hyning, Mason A. Jones, Rachael Bradley Montgomery, Elizabeth Bottner, and Evan Tansil. 2024. Information Wayfinding of Screen Reader Users: Five Personas to Expand Conceptualizations of User Experiences. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 47, 7 pages. https://doi.org/10.1145/3663548.3688539

[51] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* 1 (2012), 4–2.

[52] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What Frustrates Screen Reader Users on the Web: A Study of 100 Blind Users. 22, 3 (2007), 247–269. https://doi.org/10.1080/10447310709336964

[53] Christina Low, Emma McCamey, Cole Gleason, Patrick Carrington, Jeffrey P. Bigham, and Amy Pavel. 2019. Twitter A11y: A Browser Extension to Describe Images. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Pittsburgh PA USA, 551–553. https://doi.org/10.1145/3308561.3354629

[54] Yao Lyu, Jie Cai, Anisa Callis, Kelley Cotter, and John M. Carroll. 2024. "I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.": A Study of Blind TikTokers' Content Moderation Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 741, 15 pages. https://doi.org/10.1145/3613904.3642148

[55] Yao Lyu, Jie Cai, Bryan Dosono, Davis Yadav, and John M. Carroll. 2024. "I Upload...All Types of Different Things to Say, the World of Blindness Is More Than What They Think It Is": A Study of Blind TikTokers' Identity Work from a Flourishing Perspective. *The ACM Conference on Computer-Supported Cooperative Work And Social Computing (CSCW)* (April 2024). arXiv:2404.14305 [cs]

[56] 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 TikTokers' Infrastructuring Work to Build Independence. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW (2024). https://doi.org/10.1145/3637297

[57] Yao Lyu, Tawanna Dillahunt, Jiaying Liu, and John M. Carroll. 2026. "My Brother Is a School Principal, Earns About $80,000 Per Year... But When the Kids See Me, 'Wow, Uncle, You Have 1500 Followers on TikTok!'": A Study of Blind TikTokers' Alternative Professional Development Experiences. arXiv:2601.10956 [cs.HC]

[58] Yao Lyu, Jessica Shen, Alina Faisal, and John M. Carroll. 2026. I'm Constantly Getting Comments Like, "Oh, You're Blind. You're Like the Only Woman That I Stand a Chance With.": A Study of Blind TikTokers' Intersectional Experiences of Gender and Sexuality. arXiv:2601.10957 [cs.HC]

[59] Renkai Ma and Yubo Kou. 2022. " I'm not sure what difference is between their content and mine, other than the person itself" A Study of Fairness Perception of Content Moderation on YouTube. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[60] J. Nathan Matias. 2019. The Civic Labor of Volunteer Moderators Online. *Social Media + Society* 5, 2 (2019), 1–14. https://doi.org/10.1177/2056305119836778

[61] Samuel Mayworm, Michael Ann DeVito, Daniel Delmonaco, Hibby Thach, and Oliver L Haimson. 2024. Content moderation folk theories and perceptions of platform spirit among marginalized social media users. *ACM Transactions on Social Computing* 7, 1-4 (2024), 1–27.

[62] Aiden McGillicuddy, Jean Grégoire Bernard, and Jocelyn Cranefield. 2016. Controlling bad behavior in online communities: An examination of moderation work. *2016 International Conference on Information Systems, ICIS 2016* (2016), 1–11.

[63] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. "With most of it being pictures now, I rarely use it": Understanding Twitter's Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5506–5516. https://doi.org/10.1145/2858036.2858116

[64] Donald Moynihan, Pamela Herd, and Hope Harvey. 2015. Administrative burden: Learning, psychological, and compliance costs in citizen-state interactions. *Journal of Public Administration Research and Theory* 25, 1 (2015), 43–69.

[65] Philip M. Napoli. 2019. *Social Media and the Public Interest: Media Regulation in the Disinformation Age*. Columbia University Press, New York. https://doi.org/10.7312/napo18454

[66] Jakob Nielsen and Robert L. Mack. 1994. *Usability Inspection Methods*. John Wiley & Sons, New York, NY.

[67] Donald A. Norman. 2013. *The Design of Everyday Things*. Basic Books, New York.

[68] Martin J Riedl, Teresa K Naab, Gina M Masullo, Pablo Jost, and Marc Ziegele. 2021. Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. *Policy & Internet* 13, 3 (2021), 433–451.

[69] Sarah T. Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, New Haven, CT.

[70] Anne Spencer Ross, Xiaoyi Zhang, James Fogarty, and Jacob O. Wobbrock. 2018. Examining Image-Based Button Labeling for Accessibility in Android Apps through Large-Scale Analysis. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 119–130. https://doi.org/10.1145/3234695.3236364

[71] Andreas Savva, Helen Petrie, and Christopher Douglas Power. 2015. Comparing Concurrent and Retrospective Verbal Protocols for Blind and Sighted Users. In *Human-Computer Interaction – INTERACT 2015, Part I (Lecture Notes in Computer Science, Vol. 9296)*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). IFIP TC 13, Springer, Cham, Bamberg, Germany, 55–71. https://doi.org/10.1007/978-3-319-22701-6_5 15th IFIP TC 13 International Conference on Human–Computer Interaction (INTERACT 2015).

[72] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-computer Interaction* 2, CSCW (2018), 1–27.

[73] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[74] Joseph Seering. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 107 (Oct. 2020), 28 pages. https://doi.org/10.1145/3415178

[75] Filipo Sharevski and Aziz N Zeidieh. 2023. "I Just Didn't Notice It:" Experiences with Misinformation Warnings on Social Media amongst Users Who Are Low Vision or Blind. In *Proceedings of the 2023 New Security Paradigms Workshop* (Segovia, Spain) *(NSPW '23)*. Association for Computing Machinery, New York, NY, USA, 17–33. https://doi.org/10.1145/3633500.3633502

[76] Yunhee Shim and Shagun Jhaver. 2026. Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms. *ACM Transactions on Social Computing* (2026).

[77] Kristen Shinohara, Michael McQuaid, and Nayeri Jacobo. 2021. The Burden of Survival: How Doctoral Students in Computing Bridge the Chasm of Inaccessibility. 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 376, 13 pages. https://doi.org/10.1145/3411764.3445277

[78] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven M. Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2017. *Designing the User Interface: Strategies for Effective Human–Computer Interaction* (6 ed.). Pearson, Boston. 624 pages.

[79] Lori J. Sommerfield, Eve Hill, Susan Fahey Desmond, and Matt Ater. 2025. Digital Accessibility Under Title III of the ADA: Recent Developments and Risk Mitigation Best Practices. *Business Law Today* (Aug. 2025). https://www.americanbar.org/groups/business_law/resources/business-law-today/2025-august/digital-accessibility-under-title-iii-ada/

[80] Katta Spiel, K. Gerling, Cynthia L. Bennett, Émeline Brulé, R. Williams, J. Rode, and Jennifer Mankoff. 2020. Nothing About Us Without Us: Investigating the Role of Critical Disability Studies in HCI. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020). https://doi.org/10.1145/3334480.3375150

[81] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. 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, 1–13. https://doi.org/10.1145/3313831.3376404

[82] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 341, 14 pages. https://doi.org/10.1145/3411764.3445092

[83] C. Sum, Rahaf Alharbi, F. Spektor, Cynthia L. Bennett, Christina N. Harrington, Katta Spiel, and R. Williams. 2022. Dreaming Disability Justice in HCI. *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (2022). https://doi.org/10.1145/3491101.3503731

[84] Sarit Felicia Anais Szpiro, Shafeka Hashash, Yuhang Zhao, and Shiri Azenkot. 2016. How People with Low Vision Access Computing Devices: Understanding Challenges and Opportunities. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 171–180. https://doi.org/10.1145/2982142.2982168

[85] Meini Tang, Roberto Manduchi, Susana Chung, and Raquel Prado. 2023. Screen Magnification for Readers with Low Vision: A Study on Usability and Performance. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 23, 15 pages. https://doi.org/10.1145/3597638.3608383

[86] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 415–422. https://doi.org/10.1145/985692.985745

[87] Tiziana Terranova. 2000. Free Labor: Producing Culture for the Digital Economy. *Social Text* 18, 2 (63) (2000), 33–58. https://doi.org/10.1215/01642472-18-2_63-33

[88] David Thiel and Samantha Bradshaw. 2022. *Accessibility for Trust and Safety Flows.* Stanford Cyber Policy Center. https://cyber.fsi.stanford.edu/io/news/accessibility-trust-and-safety-flows Accessed 2025-05-02.

[89] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185, 4157 (1974), 1124–1131.

[90] Twitter. 2020. Making Twitter More Accessible. Twitter Blog. https://blog.x.com/en_us/topics/company/2020/making-twitter-more-accessible Published September 2, 2020. Accessed September 15, 2026.

[91] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. " At the End of the Day Facebook Does What ItWants" How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on human-computer interaction* 4, CSCW2 (2020), 1–22.

[92] G. Vera-Amaro, R. Vera-Amaro, M. F. Mata-Rivera, and J. R. Rojano-Cáceres. 2026. Artificial Intelligence in Web Accessibility: Towards a Theory of LLM-Assisted Remediation for Visual Disabilities. *Technologies* 14, 5 (2026), 287.

[93] Emily A. Vogels. 2021. The State of Online Harassment. Pew Research Center. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

[94] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559

[95] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How Blind People Interact with Visual Content on Social Networking Services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1584–1595. https://doi.org/10.1145/2818048.2820013

[96] W3C Web Accessibility Initiative (WAI). 2023. *Designing for Web Accessibility – Tips for Getting Started.* https://www.w3.org/WAI/design-develop/design-for-all/ Accessed 8 May 2025.

[97] W3C Web Accessibility Initiative (WAI). 2025. *Understanding Success Criterion 3.2.4: Consistent Identification.* World Wide Web Consortium (W3C). https://www.w3.org/WAI/WCAG21/Understanding/consistent-identification.html Developed by the Accessibility Guidelines Working Group (AG WG).

[98] Emily Q. Wang and Anne Marie Piper. 2022. The Invisible Labor of Access in Academic Writing Practices: A Case Analysis with Dyslexic Adults. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1, Article 120 (April 2022), 25 pages. https://doi.org/10.1145/3512967

[99] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Franziska Roesner, and Kurt Thomas. 2023. "There's so much responsibility on users right now:" Expert Advice for Staying Safer From Hate and Harassment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[100] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Franziska Roesner, and Kurt Thomas. 2023. "There's so much responsibility on users right now:" Expert Advice for Staying Safer From Hate and Harassment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[101] Jacob O. Wobbrock, Shaun K. Kane, Krzysztof Z. Gajos, Susumu Harada, and Jon Froehlich. 2011. Ability-Based Design: Concept, Principles and Examples. *ACM Trans. Access. Comput.* 3, 3, Article 9 (April 2011), 27 pages. https://doi.org/10.1145/1952383.1952384

[102] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300390

[103] Shaomei Wu and Lada A. Adamic. 2014. Visually impaired users on an online social network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3133–3142. https://doi.org/10.1145/2556288.2557415

[104] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1180–1192. https://doi.org/10.1145/2998181.2998364

[105] Sophie Yates, Gemma Carey, Eleanor Malbon, and Jen Hargrave. 2022. 'Faceless monster, secret society': Women's experiences navigating the administrative burden of Australia's National Disability Insurance Scheme. *Health & Social Care in the Community* 30, 5 (2022), e2308–e2317. https://doi.org/10.1111/hsc.13669

Fig. 7. Examples of stimuli and its corresponding placeholder to instigate safety actions on X.

[106] Alice Qian Zhang, Kaitlin Montague, and Shagun Jhaver. 2023. Cleaning up the streets: Understanding motivations, mental models, and concerns of users flagging social media posts. *arXiv preprint arXiv:2309.06688* (2023).

A Appendix: Supplementary Materials

A.1 Interview Protocol

Preparation

(1) **Set up the Environment:**
- Communicate to the participant (before setting up the interview) that they should use a mobile device/desktop they are most accustomed to and a screen reader of their choice.
- Log in to the social media platform for testing, based on the user's preference (Twitter or Facebook). Use already set-up credentials for logging in.
- Ensure recording devices are ready for audio interactions.
- Have consent forms and information sheets ready for participants.

Introduction

(1) **Welcome and Consent:**
- Greet the participant, and thank them for their participation.
- Explain the purpose of the study and how their input will contribute to our research on improving social media accessibility. Also, explain to them what content moderation means, and the distinction between platform moderation and user-enacted moderation.
- Obtain informed consent (if you haven't already sent it ahead of time), ensuring the participant understands their rights and the confidentiality of their responses.

(2) **Overview of the Study:**
- Provide a brief overview of the study's structure, including the think-aloud protocol.
- Explain that they will perform specific social media tasks using their screen reader, focusing on flagging and self-moderation tools.

Fig. 8. How administrative burden accumulates across Task 1: reporting a hateful post. We organize the overarching issues into different themes. For each theme, we describe the prominent issues users faced across both Facebook and X and the letters in the small circles mark the *administrative burden of safety work* it imposes, *learning* (L), *compliance* (C) and *psychological* (P).

- Explain that you will be using a test account for this session. This is done to give participants greater flexibility with how they interact during the sessions. Their activities during this session will have no effect on their own social media accounts.
  - "We use dummy test accounts with placeholder text purporting to contain misinformation and other kinds of representative social harms instead of having actual misinformation content."

(3) **Training Session:**

- Briefly train the participants on the think-aloud method if they are unfamiliar. Encourage them to verbalize their thoughts, decisions, and difficulties through an example.
  - "Go through the posts on this feed one at a time. Realize who sent it and acknowledge the content of the post. Verbalize all your thoughts aloud as you go through the different messages."

Task 1: Initial Exploration and Discoverability

Engage user (warm-up):

- How often do you use [platform name]?
- What purposes do you use it for?
- What groups/[Community spaces] are you a part of? What kinds of messages do you see on it?
- How long have you been using it?
- Are you well aware of the different features of the platform you can use to protect yourself from harmful people or messages? Can you name a few?
- Have you ever used [tools mentioned in the previous response] and what were the circumstances that motivated you to use it?

Task Prompt:

- "Imagine you are Jenny Chang and this is your Twitter [or Facebook] feed, and you don't know the person Marcus Chang. There are some posts that purport to contain hateful speech or misinformation through placeholder text. Scroll through the media feed and when you see something disturbing, report it. Please describe out loud everything you are doing and experiencing as you work through the process."

Task: Navigation

(1) So, you've decided you want to report a post. What icons/texts do you think you can use to initiate the reporting process? How do you feel about the placement of these icons? How could this be designed better?
(2) What did you understand happened after you clicked on the report icon/button? Is it clear how you are expected to proceed?
(3) What did you think of the use of multiple steps within this flagging interface? Was anything confusing? How could this interface be made more accessible for visually impaired users who are new to the platform?

Task 2: Feedback and Flexibility

Task Prompt: (For the already flagged piece previously)

- "How clear was the feedback you received after flagging the content? Is any information missing? Were you able to understand the outcomes of your actions? What additional information would you have liked to see about how platforms would review your flag? How could this have been designed better?"

Follow-up Prompt (for Task 1):

- "What made you choose [specific option they chose]? Did you know what you were going for? Were you aware of the existing categories or were you exploring the options for flagging?"

Task 3: Consistency and Shortcuts

Task Prompt:

- "Check if there is any hateful post made inside a community [group on Facebook] and try to report that community. Describe each step you take. Verbalize all your thoughts aloud."

Follow-up Prompt:

- Retrospective Verbal Protocol: "Did you notice any consistency issues in the placement of flagging features across different sections? Do you have any specific suggestions or feedback?"

Task 4: Blocking

Task Prompt:

(1) "Block the user who has posted harassing content so they won't be able to reach you again. Describe each step and any obstacles you face. Verbalize all your thoughts aloud."
(2) What do you think of the information platforms gave after you finished blocking?

Follow-up Prompt:

- How difficult was it to figure out which steps to take to go to the privacy settings page? How could these settings be made more easily discoverable and navigable?
- "How accessible and easy were the privacy settings to adjust? What challenges did you encounter while blocking a user?"

Task 5: Messaging/Privacy Settings and Unblocking

Task Prompt:

(1) Navigate to the settings and adjust who can see your posts. This functionality is available on X [Facebook]. Describe each step and any obstacles you face. Verbalize all your thoughts aloud.
(2) "Adjust your account settings to prevent receiving messages from non-friends."
(3) "Then attempt to unblock a user you previously blocked. Describe the process and any issues you encounter."

Follow-up Prompt:

- How difficult was it to figure out which steps to take to arrive at the message settings page? How could these settings be made more easily discoverable and navigable?
- "What was it like to adjust the messaging settings?"
- "How flexible and accessible was the unblocking process?"

Task 6: Sensitive Content Settings

Task Prompt:

- "Did you see a post on this feed which was political? Suppose this was suggested by the algorithm. Can you see if there is a way to navigate to content settings and toggle them on or off to see less political or sensitive content? Describe each step and any difficulties you encounter (if not already previously answered)."

Follow-up Prompt:

- "How discoverable and navigable were the sensitive content settings? How easy was it to use the word filters to block specific content?"

Task 7: Gathering User Feedback

Task Prompt:

- "Reflect on your experience with the self-moderation tools we explored today. If you encounter something inappropriate during your social media use, what features, if any, would you use? Why? What improvements would you suggest?"

Closing

(1) Thank the participants for their time and valuable input.
(2) Provide information on the next steps, including how the data will be used and when they can expect any follow-up communication.
(3) If applicable, provide any compensation or incentives promised to the participant.