

Cleaning Up the Streets: Understanding Motivations, Folk Theories, and Concerns of Users Flagging Social Media Content

ALICE QIAN, Carnegie Mellon University, USA

KAITLIN MONTAGUE, Rutgers University, USA

SHAGUN JHAVER, Rutgers University, USA

Social media platforms offer flagging, a technical feature that empowers users to report inappropriate posts or bad actors to reduce online harm. This feature, available now on all major social media platforms, is enacted through deceptively simple interfaces that disguise complex underlying interactions among users, algorithms, and moderators. Through interviewing 25 social media users with prior flagging experience, we examine end-users' understanding of flagging procedures, explore the factors that motivate them to flag, and surface their cognitive and privacy concerns. We found that a lack of procedural transparency in flagging mechanisms creates gaps in users' folk theories about how the mechanisms operate, yet users recognize the importance of platforms providing flagging options. Our analysis highlights how flags raise critical questions about distributing labor and responsibility for addressing online harm between platforms and users. We recommend innovations in the flagging design space that enhance user comprehension, ensure privacy, and reduce cognitive burdens.

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

Additional Key Words and Phrases: content moderation, flags, transparency

ACM Reference Format:

Alice Qian, Kaitlin Montague, and Shagun Jhaver. 2026. Cleaning Up the Streets: Understanding Motivations, Folk Theories, and Concerns of Users Flagging Social Media Content. 1, 1 (October 2026), 28 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

End-users have several options to take action when encountering inappropriate posts on social media platforms, from regular interactions built into a platform (e.g., replying to the post, unfolding the poster) to content moderation-specific interactions. Of these options, only through **flagging** can end-users directly request platform administrators to take site-wide actions against any content. Platforms maintain content regulation systems, comprising a coordinated deployment of automated tools and human reviewers [34, 41], that regularly review flagged items and trigger sanctions, such as removing flagged posts or banning flagged accounts [32, 35]. Thus, the *flag* potentially serves as a powerful and empowering tool for user-driven moderation.

*This work was conducted while the author was a student at the University of Minnesota.

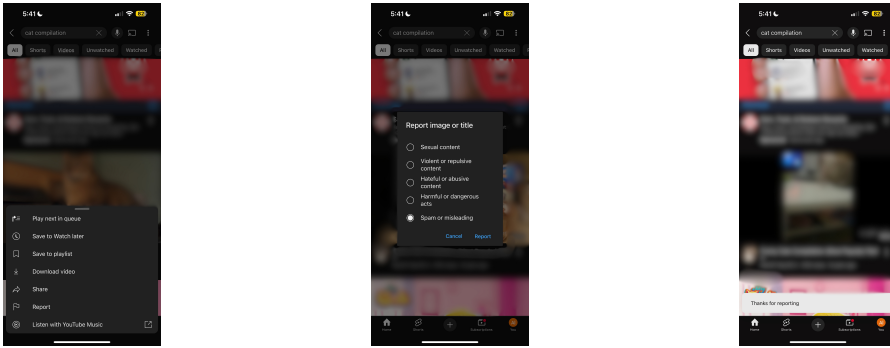
Authors' Contact Information: Alice Qian, aqzhang@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Kaitlin Montague, kaitlin.montague@rutgers.edu, Rutgers University, New Brunswick, NJ, USA; Shagun Jhaver, shagun.jhaver@rutgers.edu, Rutgers University, New Brunswick, NJ, USA.

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 the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM XXXX-XXXX/2026/10-ART

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>



(a) The report button for flagging content on YouTube.

(b) Categories for flagging content on YouTube.

(c) “Thank you” pop-up after users flag content on YouTube.

Fig. 1. YouTube’s flagging interface pages.

Though at first glance, flag implementations may signal platforms’ commitment to enacting democratic governance and fostering user communities, in practice, flags are often designed without adequately considering the needs of users who rely on them. As Crawford and Gillespie describe, flagging is a *“complex interplay between users and platforms, humans and algorithms, and the social norms and regulatory structures of social media”* [17]. While their work theorized the conceptual significance of flags as critical socio-technical mechanisms, there remains a need for more empirical research examining *how* users—the primary stakeholders in the flagging processes—perceive flagging mechanisms and *why* they choose to engage in flagging. This lack of focus on user perspectives is particularly striking, given that flagging processes across platforms vary significantly in the input required of flag submitters and post-flagging review and notification procedures.

For instance, when a user flags a video on YouTube, they see a quick pop-up that thanks users for flagging content with no other contact from the platform about the outcome of their flag (Fig. 1). In contrast, when users flag content on Facebook, the platform typically tells them they will be notified when a decision is made. It provides them with additional safety options such as blocking, muting, and hiding the flagged content. Facebook also offers users a centralized interface where they may revisit past flags to check their review status (Fig. 2). Although such varied implementations are not inherently problematic, limited transparency about design decisions and the lack of empirical insight into user perceptions make it difficult to judge whether current flagging processes meet user needs or how they might be improved.

In this study, we analyze user experiences with flagging mechanisms on widely used social media platforms, such as Facebook, X, and TikTok, and ask three key research questions. First, prior research has shown that platforms’ opaque flagging processes may lead users to develop “folk theories” (i.e., “socially constructed conceptions of how a platform works” [18]) about moderation decisions. These folk theories raise concerns about potential censorship and silencing through practices like shadowbanning [3, 64, 70, 98]. While prior work has mostly examined users as content creators who are flagged [92], less is known about users who flag others’ content. Flagging interfaces reveal differing amounts of information about how reports are processed, review timelines, and reviewer identity, yet it is unclear how users interpret these signals or what else shapes their understanding of flagging affordances and review processes. Therefore, we ask:

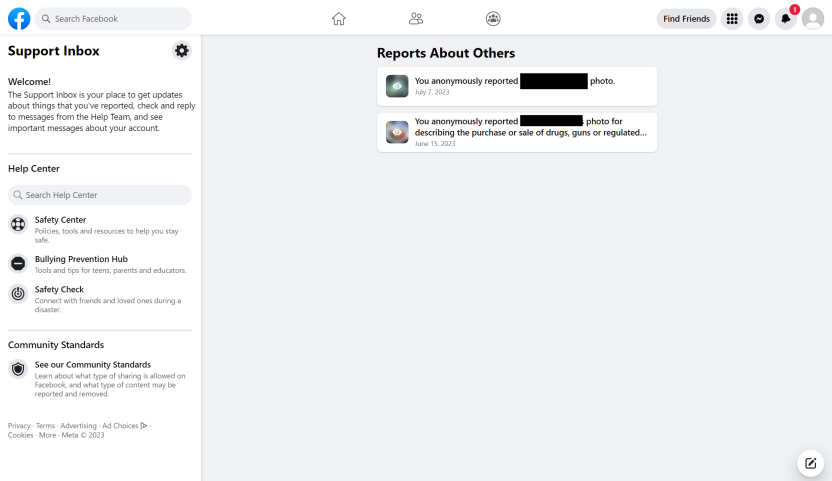


Fig. 2. Facebook's support inbox dashboard showing posts a user has flagged.

RQ1: *What are users' understandings of how social media platforms implement flagging mechanisms, and what factors influence this understanding?*

Next, we examine *why* users choose to flag content on social media platforms. While scholars have previously examined users' motivations for using personal moderation tools [49], such as word filters [45] and blocking [46], there is limited understanding of what motivates users to flag harmful content on social media platforms. Further, it remains unclear when and why active flaggers disengage from flagging. Flagging mechanisms have been studied in gaming contexts, where users may flag for reasons unrelated to platform-defined toxicity [54]. However, social media platforms serve as central hubs for public discourse, making it critical to understand the factors driving users to flag or abstain from flagging harmful content. Thus, we ask:

RQ2: *What motivates users to engage (or not engage) in flagging on social media platforms?*

Submitting a flag is a form of anonymous, voluntary labor that asks users to navigate multi-step interfaces and fit their concerns into platform-defined labels of inappropriate conduct. Prior work has documented the physical and emotional burdens of moderation on end-users and content moderators [49, 58, 75, 100] and shown that people weigh privacy risks and protections when reporting harassment on messaging platforms [96]. Yet we know little about how these cognitive and privacy burdens manifest in flagging on mainstream social media platforms. To better understand these challenges, we ask:

RQ3: *What are users' cognitive and privacy concerns about engaging in flagging on social media platforms?*

To answer these questions, we conducted semi-structured interviews with 25 social media users who had recently flagged content. Applying interpretive qualitative analysis [65] to our interview data, we organize our findings around users' understandings of flagging, their motivations for engaging with it, and their cognitive and privacy concerns. We show that the lack of transparency in flagging creates substantial gaps in users' folk theories. Given this lack of transparency, users develop incomplete and uncertain interpretations of how flags work. Yet, they develop strategies to leverage these interpretations, using flags alongside other site mechanisms, such as blocking and

private messaging, to address online harms. A belief in generalized reciprocity motivates users to flag inappropriate content—they expect a collective flagging effort would help keep online spaces clean, yet this also exposes tensions between flagging as a civic right and as an obligation that shifts unpaid digital labor onto users. We surface the reservations users have about the cognitive burdens of flagging and examine their privacy concerns when discussing potential innovations to flagging mechanisms.

Our main contributions to the HCI literature on content moderation mechanisms include:

- (1) *Empirical data on user experiences and perspectives about flagging on social media platforms.* We extend Crawford and Gillespie’s conceptualization of flagging [17] with findings organized around users’ understandings of flagging, their motivations for engaging with it, and their cognitive and privacy concerns. Our findings surface tensions between viewing flags as a right and as an obligation, and illuminate critical concerns about the labor of flagging from the perspective of flaggers.
- (2) *Design opportunities.* We provide empirically informed design recommendations for innovating the current flagging systems in three crucial areas: (1) improving literacy regarding the purpose and benefits of flags, (2) enhancing user interaction during flag category selection, and (3) offering satisfactory post-reporting feedback. We describe how our proposed innovations to flagging mechanisms can address user needs while attending to their privacy and cognitive burdens. We explore how incorporating seams [25] into flagging interfaces can help users understand their operations and ensure transparency throughout the flagging process.

2 Related Work

2.1 Content Moderation Affordances on Social Media

Most contemporary platforms hosting user-generated content offer a mechanism for users to provide feedback on the content they see that violates the platform’s guidelines. This mechanism is usually called “flag” or “report.” Platforms use flags to collect instances of potentially inappropriate content, which human moderators or bots subsequently review [54]. If this review determines that the content in question indeed violates the platform’s guidelines, the flagged user may face direct consequences, such as content removal or account suspension [40, 42].

We contextualize flags as part of a collection of mechanisms that let users express their voice in content moderation systems. Flags are often accompanied by a variety of *personal moderation tools*, which have been defined as “tools that let users configure or customize some aspects of their moderation preferences on social media” [49]. These personal moderation tools include *account-based tools* that let users mute or block specific accounts [29, 46, 63] and *content-based tools* that let users hide comments containing configured keywords [45] or set up content sensitivity controls [50]. Together, these mechanisms allow users to express dissatisfaction about norm-violating content (e.g., hate speech, violence, spam) [5, 55, 82] or behaviors (e.g., harassment, bullying, scamming) [86].

We focus on flagging, one of the most conspicuous content moderation mechanisms available to general end-users. While personal moderation tools change only the configuring user’s news feed, flags enable site-wide content regulation, provide higher granularity in letting users express their dissatisfaction, and prompt users to justify their actions. *Appeal mechanisms* [94, 95], which let moderated users report their dissatisfaction with content moderation sanctions (regarding content posted by themselves) and request that they be reversed, are often used to contest platform moderation actions on flagged content or other content removed by administrators. Users may also utilize the above moderation mechanisms in combination with regular platform interactions that

demonstrate positive or negative feedback (e.g., like and dislike buttons) [52] or other social signals (e.g., unfollowing accounts, engaging in counterspeech) [80].

Users with special roles (e.g., community moderators) usually have access to a wider range of moderation tools [37, 89]. For example, Reddit moderators may independently program Automod, a bot, to automatically identify and take action against content that infringes upon their community's guidelines [41, 101]. Similarly, Twitch streamers and YouTube creators rely on third-party applications for harmful content detection [12, 45]. We do not focus on such specialized tools because this study is examining regular end-users who lack access to them. However, we draw from literature on accountability and division of labor involved in the deployment of such tools [2, 14, 41] to examine how end users perceive the use of automation and human moderators to review flagged content.

2.2 User Perspectives on Content Moderation

Prior work has emphasized the importance of attending to user perspectives to inform platform practices and policymaking, especially since content moderation practices can shape public discourse [56, 72]. Inspired by these calls, our research on users' perceptions of flagging mechanisms highlights existing user needs to inform changes to how platforms design and administer these systems. Discussions of user perspectives often center on who owns the responsibility for platform governance. Gillespie argues for a collective civic responsibility to govern platforms [31], yet specific stakeholder roles remain unclear. Prior research shows that users tend to attribute responsibility for detecting and intervening on inappropriate content to platforms [72] or human moderators [67] rather than themselves. However, Jhaver and Zhang [50] found that users prefer personally configurable moderation tools to platform-enacted bans of hate speech, violent content, and sexually explicit content. On the whole, this literature suggests that users expect platforms to offer a baseline level of safety from online harms, but still desire to further shape their feeds themselves. We build on this work by examining how end-users perceive their responsibilities to flag, fitting in with how they assume the obligations to regulate inappropriate content are divided among all stakeholders.

Prior research also documents the physical and emotional burdens that moderation places on regular users [49], volunteer moderators [23, 41, 58, 78, 100], and commercial moderators [76, 88], as well as tensions around how volunteer labor contributes to platform profit [59, 100]. While emotional burdens on volunteer and hired moderators have been examined in detail [23, 76], little prior work has considered the additional burden on regular users who report content. Our study addresses this gap.

People with marginalized identities are more likely to experience harm on online platforms [3, 24]. For example, Black and female identifying users disproportionately experience a range of unwanted behaviors such as hate speech and doxxing [66]. To address these disparities, researchers have prioritized the needs and perspectives of vulnerable groups [3, 36, 60, 92]. For instance, Blackwell et al. examined HeartMob, a platform built for people facing severe online abuse, and argued that design and moderation must integrate vulnerable users' unique needs to fully address harassment [8]. This motivates our focus on diverse users, including those from marginalized groups, and on fears of retaliation, privacy concerns, disproportionate labor, and apprehensions about misuse of flagging mechanisms.

Finally, when examining the design of moderation systems, it is also important to consider end-users' preferences for control and transparency. Most social media platforms do not offer adequate visibility into the rules and decision processes behind moderation decisions [91], and interfaces often obscure what is done by automated tools versus human reviewers [51, 68]. Some researchers have examined how different configurations of moderation tool design shape the control that users perceive over their social media feeds [45, 49]. Additionally, prior research shows that

when moderation systems indicate the reasoning behind content regulation decisions, user attitude toward those decisions may improve, resulting in higher-quality user posts [40, 43, 53] or generally increased trust [10]. Building on this literature, we examine users' transparency needs regarding flagging and offer empirical insights about how they seek individual agency in flagging processes.

2.3 Flagging Mechanisms on Social Media Platforms

While it is unclear when flags were first implemented on social media platforms, the notion of raising a concern to a community leader or administrator has long been used to express dissatisfaction civilly [21]. Flags are a crucial means for users to express such dissatisfaction in online environments. In this section, we provide an overview of prior research on flagging to better situate the context of our study.

Previous research examined flagging from end-users' perspectives as creators of sanctioned content, especially in the context of how users develop "folk theories" [19] about moderation decisions when platforms fail to provide adequate or any explanations for those decisions [92]. The opaqueness of flagging has led to speculations about platforms making controversial decisions about flagged content without the flagged user's knowledge, such as with shadowbanning (where platforms hide a user's content from others' feeds without removing it entirely) [64, 98]. This has raised concerns about the use of flagging to silence [3] or otherwise censor users [70]. Our study expands on this work by examining users' perspectives when they are flagging others' content rather than being flagged themselves.

The definition of how "flaggable" a post is, or its "flaggability," for users may differ from platforms' definitions of what is inappropriate. For example, users on gaming platforms utilize flags to report what they defined as toxic content but also flag teammates to whom they attributed game losses [54]. Zhao and Chen [102] presented a case study of mass reporting on Weibo where fans of Chinese celebrities publicly coordinated to report content critical of those celebrities in an effort to take it down. Prior scholarship has also documented that users differ in what content they interpret as harmful enough to flag [44, 61], and platforms have differing guidelines on what is deemed inappropriate and thus flaggable [69]. However, there exists a gap in our understanding of how users navigate conflicts between their values and platform guidelines for flagging. Thus, we examine the decision-making process behind how users choose to flag (either appropriate or inappropriate) content online. Drawing from prior literature, we refer to inappropriate content as a multitude of undesirable behaviors [82] that include but are not limited to trolling [30, 71], invasion of privacy [1, 9, 16, 79], public shaming [33, 77, 95], and interpersonal harm [73, 96].

From end-users' point of view, engaging in flagging involves a range of emotions and safety considerations. Wang et al. [96] examined users' privacy concerns when reporting messages or accounts perpetrating harassment on end-to-end encrypted messaging platforms. They found that users make nuanced decisions about whether and how much to report based on their perceived trade-offs between privacy risks and protections. Further, these decisions are influenced by their trust in messaging platforms and community moderators. We build on this work to examine social media users' privacy concerns when they consider reporting inappropriate content. Riedl and Newell [73] reported a case study of Taylor Swift fans reporting nonconsensual pornographic deepfakes of the pop superstar on X. They noted that such reporting is motivated by fans' strong sense of responsibility to protect Taylor Swift online. We add to this research by examining how users perceive their versus platforms' flagging obligations more broadly.

3 Methods

3.1 Participant Recruitment

We recruited participants by posting on X, NextDoor, and Craigslist. In selecting these platforms to place our recruitment calls on, we aimed to reach out to a diverse set of potential participants. We asked candidates to submit an online form that included questions about whether they use social media daily, which social media sites they use, whether they have encountered toxic content on social media, whether they flagged a post in the past week, and their demographic information. The IRB at Rutgers University approved this study for recruitment and data collection. We developed the above questions to recruit participants who could recall experiences where they chose to flag or refrain from flagging in detail and used these experiences to ground our interviews, enabling more specific feedback on the purpose and operation of flags, motivations for flagging, and concerns about the process.¹ We included an option to attach a screenshot of a post the user recently flagged. We also posed an open-ended question: “What is your perspective on how social media platforms can improve flagging mechanisms?”

We screened all submissions for thoughtful and legitimate responses, and although our recruitment strategy targeted a diverse pool of potential participants, we ultimately received an overwhelming response from individuals identifying with underrepresented groups (e.g., Black and LGBTQ+ users). This outcome is consistent with previous research, which indicates that users from marginalized communities are more likely to encounter harmful content and thus engage with content moderation mechanisms [24, 46]. Given that our study focuses on qualitative insights from individuals with recent flagging experience, our goal was not to achieve broad population generalizability, but rather to gain a deep understanding of the perspectives of active flaggers. This approach also aligns with previous content moderation studies that have oversampled individuals with marginalized identities to ensure their perspectives on experiencing and addressing online harm are robustly represented [36, 49]. Table 1 presents the demographic details of our participants. Although our recruitment spanned several countries, we note that all members of our research team were based in the United States at the time we conducted this study.

3.2 Data Collection and Analysis

In total, we collected data through 25 semi-structured interviews with social media users. Participants were compensated with \$20 USD for their time, irrespective of the interview duration. This amount was above minimum wage at the interview time. Interviews lasted between 30 and 90 minutes and were all conducted remotely over Zoom. All interviews were recorded and transcribed. At the start of each interview, we asked participants for explicit verbal consent for their audio and/or video to be recorded for us to conduct analysis. We purposely did not use any flagging interfaces as probes to avoid the focus on a single platform. Instead, we grounded our discussions in users’ prior interactions with flags on the platforms they use. We also attended to the similarities and differences in flagging experiences across platforms whenever these comparisons seemed theoretically potent. Taking this platform-agnostic approach allowed our participants to speak broadly about their experiences and concerns with flagging, which included issues with specific flagging interfaces, but also frequently went beyond them.

During our interviews, we asked participants about the platforms they used and why they used them to understand the contexts in which they perform flagging. Next, we requested that

¹Note that while our participant sample consisted entirely of recent flaggers (rather than non-flaggers), even these flaggers presumably may not flag every objectionable post. As such, we sought to gather insights from them regarding factors that not just motivate them to flag but also demotivate their flagging (RQ2). Still, the never-flaggers’ motivations may present additional nuances and require an independent investigation.

#	Age	Gender	Race	Occupation	Country	Social Media Platforms Used
P1	35	Non-Binary	White	Digital Content Producer	USA	Facebook, Instagram, X, TikTok, Snapchat
P2	32	Female	Black	Fashion Designer	USA	Facebook, Instagram, X
P3	53	Female	White	Professor	USA	Facebook, X, Reddit
P4	27	Female	White	PhD Researcher	France	X, WhatsApp
P5	33	Non-Binary	Native American	Editor	USA	Facebook, Instagram, X, Tumblr
P6	22	Female	Black	Waitress	Canada	Facebook
P7	24	Male	Black	Artist	USA	Facebook, Instagram
P8	27	Male	Black	Carpenter	USA	Facebook, X, Snapchat
P9	22	Male	Black	Fashion Designer	USA	Facebook, X
P10	30	Male	Black	Electrician	USA	Facebook, Instagram
P11	26	Male	Black	Plumber	USA	Instagram, TikTok, Snapchat
P12	20	Female	Asian	Student	Canada	Facebook, Instagram, YouTube, Reddit
P13	24	Female	Black	Student	USA	Instagram, X
P14	27	Male	Black	Uber Driver	The Netherlands	Facebook, Instagram, X
P15	26	Male	Black	Sales Representative	Belgium	Facebook, Instagram, Pinterest
P16	35	Male	Black	Engineer	UK	Facebook, Instagram, WhatsApp, X
P17	27	Male	Black	Accountant	UK	Facebook, X, Reddit
P18	25	Male	Black	Sales Agent	USA	Facebook, X
P19	25	Male	Black	Manager	USA	X
P20	34	Female	Black	Nurse	USA	Instagram, X, TikTok
P21	26	Female	Black	Assistant	USA	Facebook, Instagram, X
P22	25	Male	Middle Eastern	Recent Graduate	USA	Facebook, Reddit, LinkedIn
P23	22	Female	Asian	Student	USA	Instagram, Facebook, Reddit
P24	21	Male	Asian	Student	USA	Instagram, X, Reddit
P25	24	Female	White	Unemployed	USA	X, Instagram

Table 1. Participants’ Demographic and Platform Use Information

participants describe at least one instance of flagging content and explain their understanding of how flagging mechanisms work. We also inquired about instances when participants hesitated to flag. To delve deeper into participants’ perceptions and potential concerns, we introduced a set of hypothetical scenarios that described specific modifications to flagging designs. These scenarios were only verbally described in discussion with participants so as to inspire participant ideas without constraining them to one specific interface. For example, we asked participants to consider a scenario in which the visibility of their flagging activity was altered—such as making flags visible to friends—and to discuss the implications of such a change. These discussions were intended to elicit participant perspectives on the implications of further exploring or potentially implementing such designs and to open up the flagging design space by encouraging ideation beyond current popular implementations. Lastly, we explored participants’ experiences with the cognitive load involved in flagging, aiming to uncover any challenges or frustrations associated with the process. While we did not focus on a specific platform during recruitment, our participants often discussed flagging experiences on widely used platforms, i.e., platforms listed in the top 10 most popular social media platforms.² Specific details about which platforms each participant frequented are provided in Table 1.

We read and uploaded interview transcripts to NVivo,³ a cross-platform app for qualitative research. We applied interpretive qualitative analysis to all interview transcripts [65], reading and coding each interview soon after it was conducted, and adopting an experiential approach to data

²<https://www.pewresearch.org/internet/fact-sheet/social-media/>
³<https://lumivero.com/>

interpretation. The first two authors “open-coded” [15] interviews in multiple iterative rounds, beginning on a line-by-line basis with codes sticking close to the transcript data, and prioritizing semantic (over latent) codes. This first round of coding generated codes such as “wanting to understand platforms’ timeline for taking action with flags” and “platforms should prioritize by quantity of flags.” In subsequent rounds, we merged related codes to generate higher-level themes (e.g., a need for platform transparency about flagging processes and the necessity of flags) and identified connections between themes. Throughout coding, all the authors engaged in regular discussions and memo writing, which helped us achieve deeper reflections and stay alert to emerging themes. We treated coding as a collaborative analytic process aimed at developing a shared interpretation grounded in our data. After we finished processing twenty interviews, we reassessed our data and conducted five more interviews to better flesh out some themes, such as the “value of flagging” and “responsibility for flagging.” We stopped recruiting when ongoing analysis suggested that additional interviews were no longer yielding new themes, indicating data saturation [13, 28].

4 Findings

In this section, we begin by exploring how our participants learned about flagging and their understanding of its mechanisms (RQ1). Next, we discuss key motivations for flagging, such as preventing harm, social pressure, and civic duty, and how participants navigated tensions between flagging as civic empowerment and as an obligation that shifts responsibilities to users, as well as factors that discourage flagging (RQ2). Finally, we report participants’ concerns about the burdens of flagging and potential privacy risks (RQ3).

4.1 RQ1: What are users’ understandings of how social media platforms implement flagging mechanisms, and what factors influence this understanding?

While participants shared a common understanding of the main steps involved in flagging content, they had varied interpretations of how to approach flagging due to ambiguity in platform categories and their personal preferences. Participants’ descriptions of their interactions with flags usually noted three temporally discrete steps: determining if content should be flagged, categorizing the content and explaining why they are flagging, and receiving notifications about the outcome of their flag. P23 described her understanding of this process through her experiences on Facebook, Instagram, and Reddit:

“You stumble upon a post that you deem violates the community standards, from there platforms have a dropdown menu [that] shows the different reasons why you would report such content. Once you submit that, usually, you’d get a notification saying ‘Thank you for reporting this. We’ll let you know if this violates our community guidelines,’ and then you wait and see.”

In this subsection, we examine how participants learned that flagging exists and how they came to understand what it does, often inferring its workings from the interface and the feedback they received.

4.1.1 Discovery of the option to flag content. We found that participants became aware of the option of flagging through varied information channels rather than explicit platform instruction. This occurred in both social settings (e.g., when with friends) and while being alone (i.e., exploring site functions while on a social media platform). Some participants began flagging after watching friends flag (P9, P11, and P17). Others were introduced to the option in differing social settings: P8 and P15 were taught about flagging in school, as P15’s teacher told him: “*you could take action apart from just talking and nobody listening, you can actually flag.*” Such anecdotes show how information about

flagging often derives from offline relationships users have. Several other participants discovered flagging only when they actively explored platform options for addressing harm. These participants expressed a common sentiment that, once they noticed it, on most platforms, the user interface for flagging mechanisms made it easy for them to figure out what the function was for.

Participants also heard about the option of flagging through social media posts and news articles. For instance, they learned through posts directly on the platform (P2, P7, and P14) or from creators who mentioned flagging while discussing negative experiences with platforms moderating their content (P4 and P12). While these were the most common ways for our participants to hear about flagging, platform policies could also inform users of their choice to flag. However, only P3 revealed that she read platform policies to learn about flagging. Together, these accounts suggest that users tend to discover flagging in ad hoc ways through social ties, interface exploration, and public discourse, rather than through systematic onboarding by platforms.

4.1.2 Review process for flagged content. Our participants held fragmented and often speculative theories of how platforms review flags. Participants' understanding differed in terms of whether they believed flagged content was automatically removed, whether there was human involvement, and if so, when it occurred and at what stage in the flagging process. While a few participants believed platforms fully automated removal decisions (P2 and P7), others imagined different mixes of automation and human judgment, and many assumed that a group of people reviewed all flagged content. P19 explained: *"[platforms] have people that work on the community guidelines...[these people] determine if posts should be taken down or not. [They] investigate such posts."* While some participants shared P19's sentiment (e.g., P4 and P20), P22 instead believed moderators mainly "confirm" or "quality check" automated removal decisions given the sheer volume of content.

Participants developed folk theories about how platforms prioritize reviewing flagged content. For instance, P15 stated that platforms *"prioritize some content over others...they base it off categories like severity,"* while P17 believed platforms used AI to determine review priority by looking at comments and other accounts interacting with a post. When encouraged to suggest solutions for greater transparency, participants described a system similar to package tracking, where they would receive confirmation that a flag was submitted, see its current review stage, and be notified of the final decision. As P3 shared: *"Tell me at what point there's automated decision-making. Tell me what point there are humans. Tell me what the overall policy is. Tell me the percentage. Give me data about how many things were reported."* These suggestions reflect a desire for review processes that are not only fair, but also legible in terms of which actors—human or automated—are involved at each stage.

4.1.3 Updates on flagged content from platforms. Participants experienced platform updates on flags as inconsistent, opaque, and often incomplete. After participants submitted their flags, they described a range of outcomes and an equally wide range of feedback from platforms. In successful cases, some participants reported being notified that the platform agreed with their flags and that the flagged content was removed (P4, P6, and P7). However, several participants noted that platforms often failed to communicate removal decisions, leaving it unclear whether flagged content was actually taken down; for example, P6, P12, and P17 only assumed posts were removed when they stopped seeing them in their feeds despite receiving no notification. P12 reasoned that the lack of follow-up from platforms could be due to the high volume of reports, stating that *"the comment just went away for me. It disappeared, and YouTube never followed up with me. I can't imagine they would because they must get hundreds of reports."*

Participants expressed frustration with the inconsistency of where notifications were placed on different platforms, as some platforms notified users as part of their typical notification feed, while others did so in a separate portal specific to flagging. P1 felt that *"a lot of times when I flag things,*

they get ignored." What it means for a flag to "get ignored" varied. Sometimes participants received no notification and saw no action on the content; in other cases, platforms promised follow up but never responded, as when X indicated it would update P16 about a decision but did not do so even weeks later. As a result, several participants began informally tracking their own flags, repeatedly checking whether the content was still visible because they did not trust platforms to communicate outcomes. These experiences suggest that the absence and ambiguity of feedback around flags can create an additional burden on users, who must do extra work to infer whether their reports led to any action.

4.1.4 Moderation actions performed alongside flagging. After flagging content, participants often took additional steps before or alongside reporting, revealing diverse and sometimes unconventional uses of flagging shaped by their roles, values, and perceptions of platform dynamics. P16 preferred to directly message the poster before resorting to flagging: *"What I wanted to do first was inform the person nicely. Since I didn't get a reply...it's then time I see the effect of the flag."* P15 used an even more proactive strategy informed by his activist identity. After flagging content on Facebook, he and a friend confronted members of the organization in person, securing an apology and removal of the post. For P15, flagging was just one tactic in a broader repertoire of activism, complementing offline collective action and positioning platforms as only one tool among many for combating harm. These accounts also demonstrate how preexisting social roles and personal values can shape what participants see as an appropriate sequence of actions around harm.

Others mobilized social features to amplify visibility and pressure platforms to act, using existing interaction tools to raise community awareness of inappropriate content and draw faster platform attention to their flags. For instance, P21 commented on a Facebook video depicting the sexual assault of a young woman to draw more user and platform attention to how to speak out against assault: *"the consequences of [me] speaking out are not as bad as those people tolerating the incidences [of abuse] that are happening."* Similarly, P19 'liked' offensive posts because he believed they would go viral, so others would flag them, hoping this would pressure the platform to review his reports faster. At the same time, many participants described deliberate moves to withdraw once they felt they had fulfilled their responsibility by flagging. For example, P20 opted to block accounts, explaining that this action would *"prevent them from coming up in [her] For You page."*

4.1.5 Misuse of flags. Participants recognized that flags can be weaponized against others, even though none reported misusing them personally. P5 noted learning about cases where users of marginalized identities who spoke out had their accounts flagged and subsequently suspended. From these anecdotes, she became worried that flagging could be an avenue for exacerbating biases against marginalized people:

"I've had friends and colleagues who have been targeted by hateful people who all report the poster's account for spam or something else fake, and then that person gets their account suspended...so it [flags] can be weaponized against marginalized people." (P5)

P3 also recounted problematic past occurrences where groups of "trolls" on platforms such as Reddit banded together to suspend accounts they felt did not 'belong' in a certain online community. Flags in these accounts emerge as double-edged mechanisms: they can help protect and hold others accountable, yet are equally available as tools of exclusion, placing marginalized users at greater risk.

4.1.6 Concerns about platform motivations to provide flagging mechanisms. Dissatisfying experiences led some participants to question both the effectiveness of flags and platforms' motivations for offering them. For example, P4 shared that platforms have so much content to moderate that they have little choice but to outsource the work to their users: *"There is so much data there that's being*

posted, that it's impossible to have like a team that moderates everything." Similarly, P10 speculated that platforms benefit from flags by using them as feedback to improve recommendations and moderation, while P1 and P3 suspected that legal pressures and scrutiny from entities such as Congress also motivate platforms to offer flagging tools. These views fed into mixed levels of trust in whether flagging truly safeguards end-users. At one end, P20 thought that platforms offered flags as a smokescreen: *"I feel like they want us to feel like we can regulate what you see. They want to make us feel like we have an option in what we see and what we interact with. We have a say in the company."* Along this line, P16 speculated that platforms may be motivated to ignore flags that could benefit their business models, such as not taking offensive content down because keeping it up may help *"generate a lot of user traffic."* Conversely, some participants reasoned that platforms offer flags to genuinely protect and empower users, and keep the spaces they created safe (P1, P6, P12, and P18). Overall, these findings suggest that users hold both skeptical and hopeful views of platform motivations.

4.2 RQ2: What motivates users to engage (or not engage) in flagging on social media platforms?

Our participants framed flagging as a choice shaped by both internal commitments (such as ethical values and personal tolerance for harm) and external prompts (including social pressure and others drawing their attention to problematic content). Regardless of what influenced their decision, participants consistently emphasized that they made the final decision to flag on their own, highlighting a strong sense of individual responsibility and agency in whether to act, even as some described this responsibility as ambivalent, oscillating between civic empowerment to protect their communities and an obligation that shifts moderation responsibility onto users.

4.2.1 Moral outrage against inappropriate content. Moral judgment and a sense of personal wrongdoing were central to why participants flagged. When deciding whether to flag, participants leaned on their own ethical judgments rather than on platform-defined guidelines. For many, flagging reflected a clear prior belief that certain content should be removed. P10 and P19 relied on their own values to decide what was *"right or wrong"* to flag; as P19 put it, *"as long as it's up to my conscience that this is right or this is not right...I flag because, according to me, this post is bad."*

In practice, flagging served as a way to align participants' objections with the categories provided in the platforms' flagging interfaces, rather than clarifying their own judgments. Few participants consulted community guidelines or platform policies (P3 and P15), instead opting to select categories that best matched their personal judgments. Commonly selected categories included sexual harassment (P16, P18, P21), domestic abuse (P19), impersonation (P24), spam (P12, P16, P20), nudity (P20), racism (P2, P3, P15), and sexism (P1, P25). These choices reflected participants' ability to identify harmful posts using platform-imposed categories. Despite this, participants were often uncertain about how platforms interpreted their categorizations, especially in ambiguous cases. For instance, P20 wondered how platforms distinguish between harmful content classifications, such as teasing versus outright harassment, when reviewing flags to decide what content to remove.

4.2.2 Demands of social pressure. Prompts from friends and communities often drew participants into flagging decisions, whether through others' comments, direct messages, or offline conversations. P1 said, *"So my friend had messaged us in the group thread and was like, hey, this is really bothering me. Can you all please report and join me?"* Participants responded to these requests in different ways: P6, P15, and P16 flagged content when contacted directly by people they knew, while P12 did so only when she agreed with others' rationale. P19 was less strict when deciding whether to flag, claiming, *"I might flag the post without even checking it"* when someone asks him to flag something. Thus, participants showed varying degrees of compliance with social pressure in their

flagging behaviors. Requests from communities that shared participants' identities were especially compelling, motivating some participants to flag content when requested by others who shared their identity characteristics. For example, P5, who identified as Native American and Non-Binary, said: *"I have been encouraged particularly by other natives and other queer people and trans people to go ahead and report something."* Even when others prompted them to flag harmful content, most participants agreed to examine it but emphasized that the final decision to flag was theirs alone. Overall, requests from friends and communities shaped which posts participants considered, while the decision to flag remained something they treated as their own responsibility.

4.2.3 Collective duty of upholding community norms. Flagging was often framed as a shared responsibility: participants saw platforms as responsible for detecting harmful content, but also believed individual users should help by flagging inappropriate content. More than half of our participants believed that every user is responsible for flagging norm violations. This belief was strongly tied to their reasoning to flag content to protect others. Several participants compared the act of flagging as analogous to the collective duty of cleaning public spaces. P22 said:

"If everyone does it, then the world would be a better place...If there are some plastic bottles in the street and someone picks them up and puts them in the recycling, it's not going to remove all the plastic from the environment, save the turtles, whatever. But it's a start, and everyone should do their part no matter how small it is." (P22)

Moreover, some participants described flagging as part of a norm of mutual care. P9 recognized a norm of generalized reciprocity [99]—he flagged content to protect others with the expectation that he would receive the same protection. This sense of responsibility was often guided by values of protecting loved ones. As P20 explained, *"It's coming to protection. I'm protecting my kids, those people that I know."* This sense of responsibility extended to the broader community, as participants flagged content they saw as harmful to vulnerable groups such as children, depicting flagging as everyday work they owed both to close others and to the wider public.

4.2.4 Flagging as an avenue for user expression within content moderation systems. Among available moderation tools, participants described flagging as the only mechanism that allowed them to object to content for the entire platform, rather than simply curate what they personally saw. They valued flagging as a mechanism to express their views and offer feedback within platform moderation systems. In fact, several expressed that flagging was a "right" they had as users on social media platforms (e.g., P12). Despite not knowing the outcome of their flags or finding that flagged content was not removed, participants remained optimistic and confident about the value of flags in expressing their voices. P3 stated: *"maybe it's not really helpful, or maybe it's merely making me feel better. But if something is done about the fake posts...then I'm probably helping others avoid potentially very dangerous information"*. As such, flagging was not just a complaint mechanism, but a way to draw a line around what should be allowed on the platform and to help protect other users from harm.

4.2.5 Inadequate platform responses to users' flagging efforts. As mentioned in sec. 4.1.3, in many cases, participants felt that platforms' responses did not match the effort they put into flagging. This mismatch gradually undermined their willingness to report content. Without any feedback from platforms, participants created their own folk theories about the effects of flagging. P4 suspected that X automatically blocks flagged accounts so their posts no longer appear in her feed, which left her feeling blindsided about the outcome of her reports: *"I don't know if it's working well...I have no idea what happened to that account...I have no means to actually see the status of my reporting."* These experiences also reshaped what participants considered flaggable, leading some to equate that with content they were confident the platform would actually remove. Several participants developed a

grimmer, fatalistic take on flagging, expecting platforms to ignore most flags (e.g., P2, P3, and P9). As such, we find that when platforms provide little feedback or visible action, participants may become less willing to flag, only report cases they think might succeed, and fill in the gaps with their own theories about how moderation works.

4.2.6 Strategies to flag posts versus accounts. Participants drew a clear line between reporting individual posts and reporting entire accounts. They treated account-level reports as a higher-stakes action that demanded stronger justification, since it could lead to suspension or a ban (P16, P19, and P20). In fact, participants used specific strategies to determine when they should flag a post versus an account. Many flagged an account (as opposed to just individual posts) when they encountered extremely inappropriate singular utterances from that account and/or when they saw that account repeatedly posting inappropriate content. As P19 said: *“if it’s racist content, I’ll definitely report the content **and** the owner.”* In contrast, P20 reported an account whenever she realized she was flagging too many inappropriate posts by that account. Some participants also tried to infer the intentions behind seemingly inappropriate posts. P19, for example, said that he did not flag an account posting ostensibly controversial content because he believed that account was raising awareness about sexism toward female athletes, but he flagged a post depicting spouse abuse immediately because he perceived bad intentions on the part of the poster.

4.3 RQ3: What are users’ cognitive and privacy concerns about engaging in flagging on social media platforms?

In this subsection, we examine participants’ concerns about the cognitive effort and privacy risks associated with flagging. Cognitive load was a recurring issue, especially for participants with marginalized identities who felt repeatedly exposed to harmful content and felt responsible for reporting it. We also present participants’ ideas for how platforms could lessen this burden, such as prioritizing certain flags for review. Finally, we describe their privacy worries, including fears of retaliation from flagged accounts, and uncertainty about how visible their reports are.

4.3.1 Identity-based harms and cognitive load of flagging. Participants were distinctly aware of how flagging as a mechanism places the burden of moderation on them. Those with underrepresented identities in our sample especially emphasized the additional burden of flagging they face. For example, P1, who is non-binary, observed that they end up seeing more harmful content due to their identity characteristics: *“I will flag things because, unfortunately, if you’re a queer person, if you’re a trans person, if you’re a person of color, if you’re a woman, you will regularly encounter really heinous comments.”*

Several participants noted that flagging takes more effort than it should. As described in Section 4.2.1, categorizing harmful content can be confusing when the categories of inappropriate content platforms provide fail to cover all types of norm violations. This burden is exacerbated by platforms’ differing flagging policies, resulting in different categorizations. As P25 proposes: *“I wish you could just click a button, report it, and it would just go to whatever algorithm or software that determines this does or doesn’t meet our requirements [and] maybe you could have the option to comment.”* The cognitive load of flagging is further exacerbated when platforms fail to act on flagged content. Participants expressed frustration that their efforts often feel futile when inappropriate content remains on the platform. P12 emphasized that platforms should at least *“follow their own policy”* and be transparent in their flag review processes. Such transparency could reduce users’ uncertainty about whether their reports are taken seriously. Similarly, P16 speculated that using automated systems or bots could reduce the strain on both users and human reviewers by handling initial reviews of flagged content. These solutions highlight the importance of reducing the effort users

expend in flagging harmful content, especially for those disproportionately affected by identity-based harms.

4.3.2 Prioritization order of reviewing flags. Given the cognitive toll of flagging, particularly for users with marginalized identities who are disproportionately exposed to harm, prioritization strategies emerged as a key concern in our study. As noted in Section 4.2.5, much participant frustration stemmed from perceived delays and inadequate responses to flagged material. With platforms tasked with reviewing vast volumes of reports, improving flag review prioritization order could result in more just distributions of benefits and burdens across flag submitters.

To examine this, we prompted participants with the question: *should users be considered equally when flagging?* When presented with the idea of a “super-flagger status,” i.e., individuals given higher priority for review when they flag content, most participants said that all users should *not* be considered equally, though not all agreed on the details of that prioritization. Some participants argued that platforms should prioritize reviewing the posts flagged by influential figures such as celebrities and political leaders because their popularity suggests that many users would trust their flagging actions. P15 said: “*Basically, they have a large audience, so they should be given priority*”. Other participants argued that platforms should prioritize flags based on users’ histories and engagement. For example, they suggested deprioritizing reports from people who abuse flagging or giving more weight to flags from users who spend more time on the platform (P19). In contrast, some participants felt all flag contributors should be treated equally, invoking *equal rights for all* and arguing that having more followers should not grant extra flagging privileges. Others questioned whether prioritizing certain users’ flags was effective or fair. P20, for example, framed it as a business issue: “*I feel like we are all promoting Instagram by using it. So we are all consumers, and we are all adding revenue to the company. So they should consider us equally.*”

We further explored this topic with our participants by inquiring: *should types of flags be considered equally?* Their responses revealed sharp disagreements about how different categories of rule violations ought to be prioritized. Some argued that not all categories should be equal: P1 and P9 wanted potentially life threatening content reviewed first, while P15 placed racial discrimination at the top of the list. Others opposed any prioritization at all and argued that all flags should carry the same weight. As P18 put it: “*if something is not good for the community, it’s not good for the community.*” Still others felt that establishing a priority order was untenable; for example, P20 could not see how a system could rank racism versus sexism, given her experiences of facing both discriminations together as a Black woman. These views highlight that any attempt to prioritize flags must grapple with deep disagreements about what constitutes the most urgent harm and whose experiences should anchor those decisions.

4.3.3 Privacy for personally identifiable information. Beyond cognitive burden, participants worried about how platforms handle their personal information when they flag. They questioned where reports are sent, who can view them, and whether their identity might be exposed to the people they reported, and several said they did not want any personally identifiable information attached to their flags. P21 expressed concern about human moderators or staff at platforms seeing her information. P5 speculated that those with ill intentions could get an entry-level position and “*pass the background check*” to become a moderator. Not all participants shared these worries. For example, both P18 and P22 had few to no concerns about their personal information being leaked when they flagged content. Indeed, P18 preferred that the platform keep information about past content he flagged so that similar content would not show up in his feed in the future. These views show that while some participants see flagging data as a privacy risk and worry about insider misuse or retaliation, others are comfortable with platforms retaining this information when it clearly improves their experience.

4.3.4 Retaliation from flagged users. Although in the minority, some participants also related negative experiences where owners of accounts or content they flagged retaliated, causing them to feel unsafe. P12 developed a habit of blocking accounts after receiving hateful comments when speaking out against a post and then flagging its content: *“They (account owner of flagged content) messaged me privately, and they called me names. So I ended up having to block and then report them.”* P12’s experience shows that retaliation fears persist even with anonymous flagging and should be carefully considered before introducing features that reveal more about who flagged what.

To further explore this topic, we prompted our participants with the question of: *who should see flags?* Several participants wanted everyone to see when they flag content and to see others’ flags as well, based on a belief that the more visibly a post is flagged, the more likely it is to be taken down. Per P4: *“With only one flag...maybe moderators will not pay attention to it, but with other people flagging it means some action will be taken.”* As such, P7 added that public flags would make platforms more accountable for promptly reviewing the flagged posts. He claimed, *“If a post is being flagged, I think everybody has the right to know.”* Other participants suggested more limited visibility. P3 preferred that only a few close friends see what she flags, to avoid conflict with followers who hold *“very different views.”* P19 liked the idea of followers seeing his flags but wanted a setting to toggle this on and off, while P21 preferred that people see only some of what she flags, admitting she could be *“petty”* about certain posts. Together, these preferences point to a desire for fine-grained control over who can see flag-related activity, balancing potential benefits of visibility with ongoing fears of retaliation and social friction.

5 Discussion

5.1 The Need For and Importance of Flags

Our analysis suggests that users interact with flagging mechanisms in complex ways, with diverse ideas for improvement. Although our participants often found flagging outcomes disappointing, we uncovered a strong sentiment that flagging mechanisms are vital to their experience on social media platforms [17, 67]. Our participants described flagging as one of the few core avenues for users to express dissatisfaction with content they deemed inappropriate [48, 49]. In addition to flagging, our participants noted using personal moderation tools, such as blocking mechanisms and word filters, to protect themselves from exposure to inappropriate content [46, 49]. We found that they distinguish between flagging and personal tools like blocking: participants often flag to protect others from exposure to inappropriate posts, but they may primarily deploy blocking to protect just themselves [49]. Notably, participants often described blocking as a step taken after flagging, once they considered their reporting obligation discharged, suggesting these tools operate in sequence rather than as parallel alternatives. Thus, flagging is not an isolated tool but a key component of the sociotechnical assemblages that users rely on to address harm online. Future research should examine in greater depth how and when users deploy it alongside or in place of other moderation mechanisms and actions.

Our analysis shows that the anonymity built into flagging mechanisms allows users to avoid retaliation by not having to confront perpetrators who post inappropriate content. This extends prior findings on privacy trade-offs in reporting [93] and informal pre-reporting responses [85]: our participants often treated direct contact as the first resort and flagging as the anonymous fallback once that contact failed, positioning flagging as an escalation. In some cases, when our participants attempted to avoid flagging by commenting on an inappropriate post or messaging a user (requesting the outright removal of another user’s content), they were met with inappropriate comments or no response. They then resorted to flagging as their only option for content removal.

These insights extend our prior understanding of how users leverage commenting affordances and content moderation tools [49, 101] together to address online harms [81, 82].

We also found that flagging is a critical part of users' social media experiences because it is embedded in many social interactions users have on and offline. While prior work traces users' folk theorization of moderation to interface encounters and algorithmic experiences, our participants learned about flagging via additional means like word-of-mouth experiences [19, 26, 90] and by observing others. These examples show flagging decisions are shaped by influences beyond platform interactions.

Additionally, our findings about how our participants were often encouraged to flag in an effort to protect others or to "team up" by mass-flagging inappropriate content further demonstrate how flagging is actively employed by users for critical social functions. Notably, we found that this coordination differs in scale and intent from documented cases of mass reporting driven by fandom or censorship [60, 72, 101]; our participants describe coordination at the scale of a friend group, in which they retained individual judgment over whether to comply, pointing to a distinct, smaller-scale form of collective flagging. These social dynamics were also not limited to direct interactions with the original poster. Some participants described multi-step strategies that circumvented direct confrontation, such as contacting an organization and peers who shared similar concerns. These layered interactions illustrate how users move between their roles as targets, bystanders, advocates, and community members, and how flagging is embedded in broader constellations of care, relationships, and responsibility rather than a simple dyadic exchange between a reporter and the platform [8, 17, 73]. This suggests a productive direction for future work examining how individuals with different roles and identities coordinate responses to harm through and beyond flagging interfaces.

5.2 Platform Versus User Responsibilities for Regulating Inappropriate Content

Our analysis reveals a novel 'rights-obligation' tension in digital governance: users simultaneously demanded flagging as a democratic right while experiencing it as a moral burden. Prior work has found that users tend to locate responsibility for addressing harm in platforms [71] or human moderators [66] rather than themselves, or has called for a model of collective civic responsibility [31]; our findings show the same individual user holding both the rights framing and the obligation framing simultaneously, which means design changes that strengthen one framing, such as adding tracking dashboards to support the rights claim, may simultaneously deepen the perception of flagging as unpaid, ongoing work. Many users view flagging as a *right* that platforms should uphold, as it is the only avenue for users to request that content be removed for everyone on the platform (see Section 5.1). However, the sentiment of generalized reciprocity, along with users' inclination to take on nurturing and supportive roles in their communities [84], also leads users to view flagging as an *obligation*. This obligation becomes onerous when flagging requires extensive time and effort from end-users. Prior research has documented the cognitive burdens that engaging with other user-enacted moderation mechanisms (e.g., blocking, configuring word filters) also places on end users [29, 49, 63, 97]. Thus, design efforts to uphold users' rights in moderation systems, e.g., by enhancing control and transparency [38, 45, 91], must also address the additional obligations they may entail, especially the emotional and cognitive labor they require.

Participants also emphasized that what counts as inappropriate or high-priority harm is not uniform across users but shaped by their identities and prior experiences of abuse. For example, one Black woman in our study struggled to imagine how a platform could prioritize racist versus sexist content because she regularly encountered both kinds of attacks together. This resonates with work showing that people from marginalized groups are more likely to experience online harassment and disproportionate emotional labor in moderation processes [3, 8, 24, 36, 60, 66, 92]. When platforms

define a single priority order for reviewing flags, they risk flattening these differences. Design efforts that set review priorities should therefore attend to which communities are most affected by particular harms and involve those communities in articulating what should be treated as urgent [3, 8, 72].

Our findings suggest that users' obligations to flag often derive from their ethical judgments—our participants were driven more by a desire to protect others than to enforce platform rules. They expressed motivations rooted in broader community care, such as protecting children and family members, which sometimes led them to become activists for issues such as sexual violence. This is in line with prior research showing that individuals consider other users' vulnerabilities and safety needs when determining their own moderation preferences for a wide variety of moderation measures [39, 50, 74].

In order to ensure the potency of flagging as a user right, it is crucial that flag designs are grounded in users' motivations and needs regarding flagging rather than just platforms' needs to crowdsource [3] the detection of rule violations. As we found, flagging mechanisms often overlook the nuances of user motivations. Participants sometimes acted on their understandings of platform logic through seemingly paradoxical behaviors, such as liking posts they considered offensive in hopes of influencing how the platform would treat that content. This reflects the folk theories users develop about algorithmic visibility and moderation outcomes, and aligns with prior work on how flawed or incomplete folk theories can create unnecessary additional labor [19, 47, 92]. Platforms like YouTube⁴ prioritize users who flag with the most "accuracy," framing flagging as annotation rather than community care. This approach aligns with the metaphors of "custodians" or "janitors" applied to content moderators [31, 84] but ignores the equity and control needs of regular end-users.

In contrast, our findings liken user conceptions of flagging to a civilian picking up street trash. They may feel obligated to do so to keep their community clean and may even be encouraged by those around them, but it should ultimately be their choice to do so. We urge that improvements in flagging designs strive to align with users' model of flagging. If flagging designs are not consistent with users' motivations or do not mitigate their burdens, they risk placing disproportionate responsibility on users without support, risking the exploitation of users' unpaid labor. This concern echoes prior documentation of volunteer moderator labor and its uncompensated value [58, 99], but extends it to a population that did not choose a moderator role and receives little recognition or standing in exchange.

5.3 Design Implications for Improving Flags

Our study revealed challenges across multiple dimensions of the flagging experience. We use these findings to offer design opportunities, related considerations, and examples of feature innovations in Table 2. These opportunities foreground core values from our analysis, including transparency and user agency.

5.3.1 Flagging Literacy and Prioritization. Our analysis showed that the consistent flagging interfaces across platforms helped users locate icons and submit reports. Platforms may build on this familiarity by adopting similar iconography, placement, and terminology, while also minimizing the effort and time required to find and complete flag submissions so that flagging remains accessible even to users with limited time, attention, or technical literacy.

Some social media users may be unaware of how to use flags or not fully understand their value in addressing online harm. This lack of guidance can lead users to misuse flagging, such as by using it to simply express that they dislike content or show animosity to a user [54]. Since users often

⁴<https://blog.youtube/news-and-events/growing-our-trusted-flagger-program/>

#	Design Opportunities	Considerations	Examples of Feature Innovation
Flagging Literacy and Prioritization			
1	Enhancing flagging literacy	The purpose of flagging (i.e., to report inappropriate content) should be clearly communicated.	A pinned post with clearly defined types of inappropriate content to flag.
2	Prioritizing flags from trusted users	Further experimentation of prioritizing flags from certain users based on different metrics of positive behavior in a community should be carefully conducted.	Flags from active users with 3+ months of community membership with no violation of platform guidelines are prioritized.
3	Conveying information about flagged content	Further exploration of innovations to flags that decrease the anonymity of flags should be carefully explored while attending to users' privacy concerns.	Display that "15 other users reported this post" or personalized display of "your friend Sally Wang reported this post."
Classification and Burden Reduction			
4	Classification schemes for types of inappropriate content	Offer greater clarity about the meanings of included categories; ensure transparency in additions or changes to the classification scheme.	Examples of content that fits in each category.
5	Providing nuanced objections about flagged content	The design must balance overburdening users to include too much information (e.g., free response) and not providing enough options to indicate nuanced opinions.	Scales for severity of content (e.g., scales from 1-5 of inappropriateness).
Transparency and Post-Report Feedback			
6	Visualizing flag review status	This system should be integrated within the platform for easy accessibility.	A package-tracking-like system to showcase what review stage a flag is at.
7	Visualizing platform statistics about flagging	To preserve privacy, we may urge platforms to provide aggregate statistics about the types of content that are flagged and subsequently removed or kept.	A common dashboard or monthly report sent to users.

Table 2. Design opportunities with corresponding considerations and examples of feature innovations, organized around flagging literacy and prioritization, classification and burden reduction, and transparency and post-report feedback.

learn about flagging through peers or educators (Section 4.1.1), platforms should collaborate with schools, media sources, and content creators to promote accurate understanding and responsible use. This aligns with findings that youth consider reporting processes discouraging without adequate guidance [6], reinforcing that literacy interventions may need to reach users before their first flagging interaction rather than at the interface itself. Such efforts can help users recognize when flagging can meaningfully protect others. These educational efforts can also counter misuse, such as coordinated flagging against marginalized users [3, 64]. In this way, flag literacy becomes a principle for equitable flagging design: it supports users who are targeted by harm while making it harder for bad actors to weaponize flags. We encourage both platform-led and community-based strategies to raise awareness and deter bad actors.

Crawford and Gillespie raise the question of whether “some flags are worth more than others,” such as flagged extremist content or content flagged by a selected group of individuals [17]. Programs like YouTube’s 2016 Heroes initiative prioritized content flagged by “trusted users,” selected for their past flagging accuracy.⁴ While this system aimed to improve moderation efficiency, our findings reveal tensions between platform definitions of inappropriate content and evolving user norms. Over-reliance on accuracy overlooks the contextual, community-specific nature of harm and fails to account for flag misuse [71]. It also raises questions of equity, since giving extra weight to a narrow

set of trusted users can sideline communities that experience disproportionate harm. Community-specific moderation infrastructure has been explored for volunteer moderators [4, 14, 88], but our participants proposed competing criteria for prioritization, including audience size, platform tenure, and strict equality, with no single metric commanding consensus, suggesting that any prioritization scheme will read as illegitimate to some portion of the user base. As such, instead of a one-size-fits-all model, platforms should explore what trusted flaggers look like within specific communities. For example, highly up-voted or active community members might be well-positioned to recognize locally inappropriate content (see Section 4.2.1), and giving them more influence can align flagging systems more closely with local norms of justice and accountability.

Decisions about what flagged content should be prioritized remain unclear. Some users expected that content flagged by many people ought to receive more urgent attention and expressed interest in flag designs that display how often any content has been flagged. However, when probed about more specific disclosures such as showing who flagged the post (e.g., “your friend Bill flagged this”) or demographic breakdowns (e.g., “50% of flaggers were Asian”), they raised concerns about privacy and potential retaliation. Indeed, participants who engaged in public forms of flagging, such as commenting that a post was inappropriate before formally flagging it, sometimes faced threats in response. These findings highlight the need to approach flag transparency with careful consideration of user safety and to treat anonymity protections as a core principle of equitable flagging design, especially for users who already face disproportionate risk.

Some platforms have tested public-facing flag systems. For instance, on Zhihu, users can flag content as “controversial” and engage in follow-up discussions [57]. Yet only users with high “yan” scores—based on prior activity and platform impact—can post in these discussions, raising concerns about equity and voice. We recommend that any public or semi-public flagging features be developed with careful attention to user privacy, equity expectations, and audience management goals [62], so that the ability to contest harmful content is not limited to already powerful or highly visible users.

5.3.2 Classification and Burden Reduction. While prior research [3, 36] shows that minoritized users get disproportionately flagged, our participants expressed concern that individuals from marginalized groups face a greater burden of reporting inappropriate content. Therefore, we recommend that platforms invest in creating accurate ex-ante identification [35] and regulation of activities that harm these groups. Platforms could partner with these groups to better understand the nuances of online harms (e.g., the use of offensive slang words that outsiders may not recognize [46]) and the moderation challenges they face.

Addressing the gap between users’ perceptions of inappropriate content and platform-defined flaggable content is a critical design goal supported by our findings. Instead of a top-down taxonomy led by platform definitions, flagging interfaces could deploy bottom-up taxonomies of inappropriate content developed in collaboration with communities. Additionally, for any given taxonomy, flagging interfaces could provide examples to illustrate the meanings of content categories within that taxonomy, similar to what has been explored in the design of personal moderation mechanisms [45, 49]. Clearer descriptions through examples or other means can guide users in better identifying and reporting harmful content and ensure that definitions of harm reflect the lived experiences of communities who are disproportionately targeted. In this sense, participatory taxonomies may serve as a principle for equitable flagging design, giving affected users a greater voice in how moderation standards are articulated.

Platforms such as Facebook post generalized statements about their commitment to users, but have yet to elaborate on how user perspectives are actually taken into account. For example, Facebook’s moderation team determines whether to remove flagged content based on its Community

Standards,⁵ which it claims are “based on feedback from people and advice from experts.” However, it remains unclear how this feedback is gathered—particularly whether users with marginalized identities, who are more likely to experience harm, receive any attention. Additionally, our findings suggest that users lack a concrete understanding of how such community standards are operationalized to make content moderation decisions. Offering insights into such procedures could help platforms build trust with flag submitters and make moderation decisions more accountable, rather than appearing as opaque, top-down judgments.

No aspect of flag design should overlook the critical importance of reducing the burden of labor placed on users. Flagging not only interrupts social media use but can also be time-consuming, e.g., X requires flaggers to click through multiple options and free-response explanations as well as flag additional posts from the account they are reporting—all this to provide the platform more context for flag evaluation.⁶ For users who are repeatedly exposed to emotionally taxing content such as slurs, harassment, or graphic content, this process also demands revisiting and describing material that is upsetting or even traumatizing. Recent research reveals the psychological burdens volunteer content moderators face [83, 100]. Marginalized users in particular may shoulder this work more often, because they are more frequently targeted and feel a stronger obligation to report what they see. As our findings confirm, users view the act of flagging as a form of labor and may be less inclined to flag content if they view the task as too labor-intensive. This is in line with prior results by Bäuml et al. [6] that the German youth consider the time required for reporting as discouraging and desire a more streamlined process. Thus, we recommend that labor-intensive options that require long-form answers or the uploading of information, e.g., screenshots, are optional and implemented with a clear purpose. In fact, designing for equitable flagging during the reporting process should involve providing avenues for reporting that do not require repeatedly revisiting or elaborating on harmful material. Platforms could, for instance, borrow exposure reduction interventions from research on protecting content moderators [87, 88].

5.3.3 Transparency and Post-Report Feedback. Transparency in flagging procedures becomes increasingly critical as platforms rely more heavily on automation or related technologies (e.g., LLMs) to manage the overwhelming volume of flagged content. While users are open to the use of automation, particularly when their flags receive little attention or follow-up, they strongly prefer clarity about how automated systems and human reviewers collaborate. Our findings reveal curiosity about the division of labor within the flagging pipeline, including who determines whether content is flagged in the correct category when humans oversee automated decisions and how sanctions are applied. Greater transparency into these processes can help users better understand platform constraints, manage their expectations, and potentially reduce flag misuse. Additionally, given the greater trust users tend to place in human review, there is a need to explore how platforms may balance automation with human oversight in ways that are accountable to communities most affected by harmful content.

This emphasis on transparency extends to the broader experience of flagging. We found that at the moment, users’ lack of ability to track the flagging status on most platforms meant that even after flagging a post, they have to repeatedly check the removal status of posts they reported (see Section 4.2.5 and Section 4.2.6). Some platforms promise to notify users when an action is taken on their flags but fail to do so. Users can become disillusioned with the flagging process and doubt that platforms would take any action in these cases, a concern echoed in previous findings on platforms gaslighting or silencing users with other moderation mechanisms [3, 22]. Users’ repeated monitoring of submitted flags adds to the ongoing cognitive and emotional labor, particularly

⁵<https://transparency.meta.com/policies/community-standards/>

⁶<https://help.x.com/en/rules-and-policies/x-report-violation>

for those who already face disproportionate exposure to harmful content, and erodes trust that platforms will uphold their side of the responsibility.

We recommend that platforms institute information and visualization systems that allow users to easily monitor the status of their flags. This need for systems enhancing procedural visibility aligns with prior research emphasizing the importance of transparency in platform processes. For instance, the lack of transparency has been noted as a critical concern in content removal decisions [10, 40, 43] and broader platform moderation practices [51]. However, most platforms either lack post-flagging dashboards entirely or rely on rudimentary interfaces, with Facebook being a notable exception—though even Facebook’s system is limited in transparency and hosted separately from the main platform. We also note that YouTube’s Heroes program mentioned earlier also provided users with tools to “track their own contributions and see their overall impact.”⁷ However, access to this dashboard is limited to those who volunteer and are selected for the program. While there are complexities in other tools provided in the Heroes program (e.g., mass-reporting tool), it is unclear why YouTube or other platforms could not simply provide a dashboard to track flag status to a broader selection of users, as this program was launched several years prior to our study. Making such tools broadly available would not only acknowledge users’ unpaid moderation work but also give them clearer evidence that their reports lead to action, which is central to perceptions of fairness and accountability. Overall, we find that these examples of inconsistent implementation with a lack of evaluation of such dashboards present significant opportunities for innovation in the design of post-flagging mechanisms. We recommend that such innovations incorporate principles of equitable design, and ensure procedural visibility and user agency.

In addition to technological and design improvements, the push for transparency ties into broader calls for policy reform. Scholars have highlighted the need for regulating content moderation practices [11, 56] and explored how existing policies, such as the EU Digital Services Act, address problematic platform behaviors like shadowbanning [20, 98] and unjust moderation standards [20]. The scope of these transparency requirements varies by jurisdiction; for instance, the EU’s Digital Services Act mandates a statement of reasons for content removal decisions, while comparable requirements remain less developed in regions such as Canada and the US [27]. This suggests that improvements to flagging transparency would need to be paired with region-specific regulatory advocacy efforts. Moreover, such legal and regulatory attention should extend to flagging reforms, ensuring that platforms are held accountable for transparent and equitable content moderation and that users have enforceable rights to information about how their reports are processed and acted upon.

Finally, privacy concerns remain central to discussions about increasing transparency in flagging mechanisms. To balance these concerns with the need for transparency, we propose the design opportunity of providing visualizations for aggregated statistics about flagged content, which offers users meaningful insights while safeguarding individual privacy. For example, aggregate statistics can surface patterns in what types of content are removed and whose reports tend to be acted on, supporting public scrutiny of potential biases in enforcement without exposing individual flaggers to retaliation.

5.4 Limitations and Future Work

In this study, we conducted semi-structured interviews with 25 social media users who recently engaged in flagging. Though our data helped us examine nuanced user experiences and tensions with flagging in depth, the qualitative nature of our study design and the size of our sample

⁷We were unable to find any documentation of what the dashboard looks like aside from a screenshot of flagging forum questions.

limited us from generalizing our results and learning about the relative popularity of various design choices in flagging interfaces. Our platform-agnostic approach means our findings reflect flagging experiences across several platforms, which limits our ability to attribute participants' views to the flagging design of any single platform. Future work could compare flagging experiences across platforms to examine how specific interface and feedback designs shape users' folk theories. Relatedly, while our participants were from several countries, our data collection and analysis did not focus on yielding inter-country differences in flagging perceptions; future work could seek to address this limitation. Our data could have recall bias; studies that gather participant insights at the moment of flagging could offer additional valuable insights. We also acknowledge potential desirability bias in our interviews, as participants may have avoided discussing the misuse of flags to present themselves more favorably [7]. Moreover, our findings may be limited because online behaviors and perceptions may change over time.

Future research could investigate factors not addressed in our study, such as the influence of education and digital literacy on the folk theories of flagging. Additionally, we recruited participants with prior flagging experience on widely used social media platforms such as Facebook and X, and included those with underrepresented identities that may have further motivated them to engage in flagging. Population-wide survey studies are better suited to explore the distribution of flagging awareness levels among users and across a wider range of platforms beyond the most widely used ones. Further, longitudinal methods may provide additional insights into how factors such as the volume of inappropriate content, demographic characteristics, and the experiences users have with flagging shape their perceptions of flagging over time.

We focused on the perspectives of end-users who have engaged in flagging in this work in our effort to provide empirical data that contributes to Crawford and Gillespie's existing conceptualization of flagging [17]. It would be valuable to examine the dynamics and challenges of flag processing from the perspective of platform administrators and moderators. For example, it might be insightful to learn how false flagging is identified and sanctioned and design solutions that discourage users from submitting flags in bad faith.

6 Conclusion

At a time when user participation in social media moderation is increasingly recognized as vital [45, 49, 95], our study serves as a starting point for exploring how users report inappropriate content and how mechanisms such as flags can be designed to better facilitate these interactions. Our research examined why end-users flag, their experiences with flagging mechanisms, their folk theories of the flagging pipeline, and reasons why they may stop flagging. We found that users perceive flags not as a stopgap solution that compensates for platforms' limited moderation resources, but as an essential avenue for voicing their disagreements with curated content. Our participants' willingness to contribute their time, even going so far as to personally contact the people who post content they flag to take it down, shows users' investment in addressing content-based harms. However, we found that a lack of procedural transparency in flagging implementations undermines users' trust in the flagging process. We call for further exploration of how flagging interfaces can be better designed, with consideration of other available moderation mechanisms, potential misuse, and user needs for more transparent and effective ecosystems to address online harm.

Acknowledgments

This work was supported by an award from the National Science Foundation's Division of Information and Intelligent Systems [grant number 2329394]. We are grateful to the reviewers for their thoughtful engagement with our paper and for helping us improve it.

References

- [1] Esma Aïmeur and David Schönfeld. 2011. The ultimate invasion of privacy: Identity theft. In *2011 Ninth Annual International Conference on Privacy, Security and Trust*. IEEE, 24–31.
- [2] Ishaku Hassan Anaobi, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Damilola Ibosiola, and Gareth Tyson. 2023. Will Admins Cope? Decentralized Moderation in the Fediverse. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) (*WWW '23*). Association for Computing Machinery, New York, NY, USA, 3109–3120. doi:10.1145/3543507.3583487
- [3] Carolina Are. 2025. Flagging as a silencing tool: Exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok. *New Media & Society* 27, 6 (2025), 3577–3595.
- [4] Automoderator. 2018. Automoderator - reddit.com. <https://www.reddit.com/wiki/automoderator>
- [5] Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the fourth workshop on online abuse and harms*. 125–137.
- [6] Julian Bäuml, Helen Bader, Marc-André Kaufhold, and Christian Reuter. 2025. Towards Youth-Sensitive Hateful Content Reporting: An Inclusive Focus Group Study in Germany. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [7] Nicole Bergen and Ronald Labonté. 2020. “Everything is perfect, and we have no problems”: detecting and limiting social desirability bias in qualitative research. *Qualitative health research* 30, 5 (2020), 783–792.
- [8] Lindsay Blackwell, Jill P Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *PACMHCI* 1, CSCW (2017), 24–1.
- [9] Danah Boyd. 2008. Facebook’s privacy trainwreck: Exposure, invasion, and social convergence. *Convergence* 14, 1 (2008), 13–20.
- [10] Jens Brunk, Jana Mattern, and Dennis M Riehle. 2019. Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, Vol. 1. IEEE, 429–435.
- [11] Nicole Buckley and Joseph S Schafer. 2022. ‘Censorship-free’platforms: Evaluating content moderation policies and practices of alternative social media. (2022).
- [12] Jie Cai and Donghee Yvette Wohn. 2019. Categorizing live streaming moderation tools: An analysis of twitch. *International Journal of Interactive Communication Systems and Technologies (IJICST)* 9, 2 (2019), 36–50.
- [13] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.
- [14] 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.
- [15] Kathy Charmaz. 2006. *Constructing grounded theory: a practical guide through qualitative analysis*. doi:10.1016/j.lisr.2007.11.003 arXiv:arXiv:1011.1669v3
- [16] Subai Chen, Chenyu Gu, Juan Wei, and Mingjie Lv. 2023. Research on the influence mechanism of privacy invasion experiences with privacy protection intentions in social media contexts: Regulatory focus as the moderator. *Frontiers in Psychology* 13 (2023), 1031592.
- [17] 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. doi:10.1177/1461444814543163 arXiv:https://doi.org/10.1177/1461444814543163
- [18] Michael Ann DeVito. 2021. Adaptive folk theorization as a path to algorithmic literacy on changing platforms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–38.
- [19] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. Algorithms ruin everything: # RIPTwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3163–3174.
- [20] Ángel Díaz and Laura Hecht-Felella. 2021. Double standards in social media content moderation. *Brennan Center for Justice at New York University School of Law* (2021), 1–23.
- [21] Julian Dibbell. 1994. A rape in cyberspace or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *Ann. Surv. Am. L.* (1994), 471.
- [22] Tom Divon, Carolina Are, and Pam Briggs. 2025. Platform gaslighting: A user-centric insight into social media corporate communications of content moderation. *Platforms & Society* 2 (2025), 29768624241303109.
- [23] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300372
- [24] Maeve Duggan. 2017. Online Harassment 2017. <https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/>

- [25] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First i like it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2371–2382.
- [26] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn'T Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea, 2015) (*CHI '15*). ACM, 153–162. doi:10.1145/2702123.2702556
- [27] European Commission. 2022. Digital Services Act, Regulation (EU) 2022/2065. European Commission. <https://digital-strategy.ec.europa.eu/en/>
- [28] Sandra L Faulkner and Stormy P Trotter. 2017. Theoretical saturation. *The International encyclopedia of communication research methods* (2017), 1–2.
- [29] 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. doi:10.1080/1369118X.2016.1153700 arXiv:<https://doi.org/10.1080/1369118X.2016.1153700>
- [30] Carolin Gerlitz and Anne Helmond. 2013. The like economy: Social buttons and the data-intensive web. *New media & society* 15, 8 (2013), 1348–1365.
- [31] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- [32] Eric Goldman. 2021. Content moderation remedies. *Mich. Tech. L. Rev.* 28 (2021), 1.
- [33] Lauren M Goldman. 2015. Trending now: The use of social media websites in public shaming punishments. *Am. Crim. L. Rev.* 52 (2015), 415.
- [34] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society* 7, 1 (2020). doi:10.1177/2053951719897945
- [35] James Grimmelman. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
- [36] 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. doi:10.1145/3479610
- [37] Taylor Hatmaker. 2021. Discord buys Sentropy, which makes AI software that fights online harassment. <https://techcrunch.com/2021/07/13/discord-buys-sentropy/>
- [38] Christopher Hood and David Heald. 2006. *Transparency: The key to better governance?* Vol. 135. Oxford University Press for The British Academy.
- [39] Shagun Jhaver. 2025. Bans vs. Warning Labels: Examining Bystanders' Support for Community-wide Moderation Interventions. *ACM Transactions on Computer-Human Interaction* 32, 2 (2025), 1–30.
- [40] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 192 (Nov. 2019), 33 pages. doi:10.1145/3359294
- [41] 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. doi:10.1145/3338243
- [42] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.
- [43] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 150 (nov 2019), 27 pages. doi:10.1145/3359252
- [44] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action. *First Monday* 23, 2 (2018). <http://firstmonday.org/ojs/index.php/fm/article/view/8232>
- [45] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X. Zhang. 2022. Designing Word Filter Tools for Creator-Led Comment Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 205, 21 pages. doi:10.1145/3491102.3517505
- [46] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (March 2018), 33 pages. doi:10.1145/3185593
- [47] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems* (2018).

- [48] Shagun Jhaver and Amy Zhang. 2022. Do Users Want Platform Moderation or Individual Control? Examining the Role of Third-Person Effects and Free Speech Support in Shaping Moderation Preferences. (2022). In Preparation.
- [49] Shagun Jhaver, Alice Qian Zhang, Quanze Chen, Nikhila Natarajan, Ruotong Wang, and Amy Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *arXiv preprint arXiv:2305.10374* (2023).
- [50] 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.
- [51] Perna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit’s Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction* 4, GROUP (2020), 1–35.
- [52] M Laeeq Khan. 2017. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in human behavior* 66 (2017), 236–247.
- [53] Nancy S. Kim. 2009. Website Proprietorship and Online Harassment. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1354466
- [54] Yubo Kou and Xinning Gui. 2021. Flag and Flagability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI ’21). Association for Computing Machinery, New York, NY, USA, Article 437, 12 pages. doi:10.1145/3411764.3445279
- [55] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. *arXiv preprint arXiv:2106.04511* (2021).
- [56] Kyle Langvardt. 2017. Regulating online content moderation. *Geo. LJ* 106 (2017), 1353.
- [57] Chen Li. 2022. Flagging Controversies: The effect of flagging mechanisms on Zhihu platform. In *European Conference on Social Media*, Vol. 9. 237–246.
- [58] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All that’s happening behind the scenes: Putting the spotlight on volunteer moderator labor in reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 584–595.
- [59] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the monetary value of online volunteer work. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 596–606.
- [60] Calvin A Liang, Sean A Munson, and Julie A Kientz. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 2 (2021), 1–47.
- [61] Zhifan Luo and Muyang Li. 2024. Participatory censorship: How online fandom community facilitates authoritarian rule. *New Media & Society* 26, 7 (2024), 4236–4254.
- [62] Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133.
- [63] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. *Available at SSRN 2602018* (2015).
- [64] 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. *Trans. Soc. Comput.* 7, 1–4, Article 1 (March 2024), 27 pages. doi:10.1145/3632741
- [65] Sharan B Merriam. 2002. Introduction to Qualitative Research. *Qualitative research in practice: Examples for discussion and analysis* 1 (2002).
- [66] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. 2022. Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [67] Teresa K Naab, Anja Kalch, and Tino GK Meitz. 2018. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20, 2 (2018), 777–795.
- [68] Marie Ozanne, Aparajita Bhandari, Natalya N Bazarova, and Dominic DiFranzo. 2022. Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society* 9, 2 (2022), 20539517221115666.
- [69] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) (GROUP ’16). ACM, New York, NY, USA, 369–374. doi:10.1145/2957276.2957297

- [70] Christopher E Peterson et al. 2013. *User-generated censorship: manipulating the maps of social media*. Ph. D. Dissertation. Massachusetts Institute of Technology.
- [71] Whitney Phillips. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press.
- [72] 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.
- [73] Martin Johannes Riedl and Ariel Newell. 2024. Reporting Image-Based Sexual Violence: Deepfakes, #ProtectTaylor-Swift, and Platform Responsibility. *Available at SSRN 4919928* (2024).
- [74] Martin J Riedl, Kelsey N Whipple, and Ryan Wallace. 2022. Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others. *Information, Communication & Society* 25, 11 (2022), 1632–1649.
- [75] Sarah T. Roberts. 2014. *Behind the screen: the hidden digital labor of commercial content moderation*. Ph. D. Dissertation. University of Illinois at Urbana-Champaign. <https://www.ideals.illinois.edu/handle/2142/50401>
- [76] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.
- [77] Jon Ronson. 2015. How One Stupid Tweet Blew Up Justine Sacco's Life. http://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html?_r=0
- [78] Minna Ruckenstein and Linda Lisa Maria Turunen. 2020. Re-humanizing the platform: Content moderators and the logic of care. *New media & society* 22, 6 (2020), 1026–1042.
- [79] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztin, Elizabeth Churchill, and Sunny Consolvo. 2019. "They Don't Leave Us Alone Anywhere We Go" Gender and Digital Abuse in South Asia. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [80] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2021. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science* 4, 1 (2021), 381–402.
- [81] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 155 (nov 2018), 27 pages. doi:10.1145/3274424
- [82] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (2021), 1–33.
- [83] Angela M Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2024. Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. *New Media & Society* 26, 10 (2024), 5677–5701.
- [84] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (2022), 621–640.
- [85] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (CSCW '17). ACM, New York, NY, USA, 111–125. doi:10.1145/2998181.2998277
- [86] Frances Shaw. 2018. Beyond 'report, block, ignore': Informal responses to trolling and harassment on social media. In *The Routledge Companion to Media and Activism*. Routledge, 395–403.
- [87] Ruth Spence, Amy Harrison, Paula Bradbury, Paul Bleakley, Elena Martellozzo, and Jeffrey DeMarco. 2023. Content moderators' strategies for coping with the stress of moderating content online. *Journal of Online Trust and Safety* 1, 5 (2023).
- [88] 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. doi:10.1145/3411764.3445092
- [89] Andrea Stockinger, Svenja Schäfer, and Sophie Lecheler. 2023. Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools. *new media & society* (2023), 14614448231190901.
- [90] Nicolas P Suzor, Sarah Myers West, Tarleton Gillespie, and Jillian York. 2017. Guiding principles for the future of content moderation. (2017).
- [91] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13 (2019), 18.

- [92] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2024. (In) visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society* 26, 7 (2024), 4034–4055.
- [93] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [94] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 167 (oct 2020), 22 pages. doi:10.1145/3415238
- [95] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 318 (oct 2021), 28 pages. doi:10.1145/3476059
- [96] Leijie Wang, Ruotong Wang, Sterling Williams-Ceci, Sanketh Menda, and Amy X Zhang. 2023. "Is Reporting Worth the Sacrifice of Revealing What I've Sent?": Privacy Considerations When Reporting on End-to-End Encrypted Platforms. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. 491–508.
- [97] 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.
- [98] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* (2018).
- [99] Monica M Whitham. 2021. Generalized generosity: How the norm of generalized reciprocity bridges collective forms of social exchange. *American Sociological Review* 86, 3 (2021), 503–531.
- [100] 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*. ACM, 160.
- [101] Lucas Wright. 2022. Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator. *Social Media+ Society* 8, 1 (2022), 20563051221077020.
- [102] Andy Zhao and Zhaodi Chen. 2023. Let's report our rivals: how Chinese fandoms game content moderation to restrain opposing voices. *Journal of Quantitative Description: Digital Media* 3 (2023). doi:10.51685/jqd.2023.006