Cleaning Up the Streets: Understanding Motivations, Mental Models, 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' understandings 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' mental models of their operations, 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. 2025. Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Content. 1, 1 (September 2025), 25 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

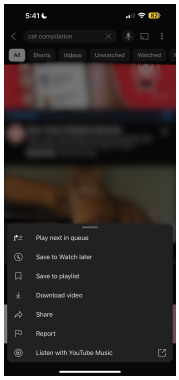
End-users have several options to take action when encountering inappropriate posts on social media platforms, from regular interactions built into a platform (e.g., replying to the post, unfollowing the poster) to content moderation-specific interactions. Of these options, it is only through **flagging** that end-users can 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 [29, 36], that regularly review flagged items and, when warranted, trigger sanctions, such as removing flagged posts or banning flagged accounts [27, 30]. Thus, the *flag* serves as a powerful and empowering tool for user-driven moderation.

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

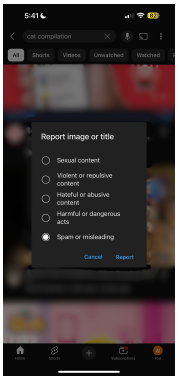
*This work was written while the author was a student at the University of Minnesota.

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.

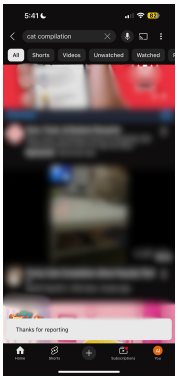
Manuscript submitted to ACM



(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.

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,"* yet they leave little room for articulation and their effects remain uncertain [15]. 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 significantly vary regarding 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 themselves may not be inherently problematic, the limited transparency regarding the design decisions and the absence of empirically backed insights into user perceptions make it challenging to assess whether current flagging processes effectively meet user needs and to identify avenues for future improvement.

In this study, we analyze user experiences with flagging mechanisms on major social media platforms and ask three key research questions. First, prior research has shown that platforms' opaque flagging processes may lead users to develop "folk theories" about moderation decisions [16], raising concerns about potential censorship and silencing through practices like shadowbanning [3, 58, 64, 87]. While much of this work has focused on users as content creators who get flagged [82], less is known about users' perspectives as audience members who flag content themselves. Flagging interfaces provide varying degrees of information regarding how user input will be processed, review timelines, and reviewer identity, but it remains unclear how users interpret these information cues and what other factors shape their understanding of flagging affordances and flag review processes. Therefore, we ask:

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

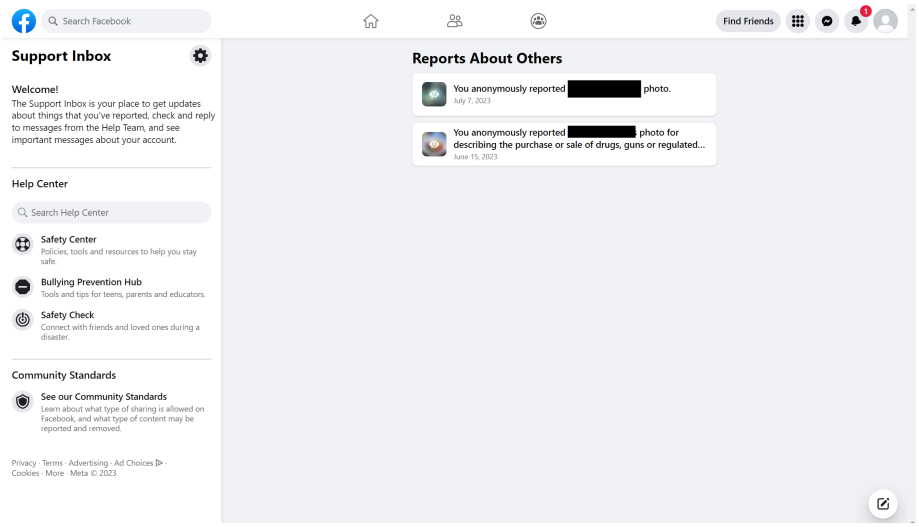


Fig. 2. Facebook's support inbox dashboard showing posts a user has flagged.

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 [42], such as word filters [40] and blocking [41], there is limited understanding of what motivates users to flag harmful content more broadly on social media platforms. Flagging mechanisms have been studied in gaming contexts, where users may flag for reasons unrelated to platform-defined toxicity [47], but 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 an act of anonymous and voluntary labor that requires users to navigate a multi-step interface, where they must categorize their concerns under the platform's pre-defined labels of inappropriate conduct. Previous research has highlighted concerns about the burdens of physical and emotional labor that moderation places on end-users [42] and content moderators [51, 69, 89]. Additionally, scholars have illustrated the privacy concerns users have when reporting harassment on messaging platforms [85], showing that users make flagging decisions based on perceived trade-offs between privacy risks and protections afforded via flagging. It remains unclear how these cognitive and privacy burdens manifest specifically in flagging contexts 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 platform content. Applying interpretive qualitative analysis [59] to our interview data, we found that flagging a post involves three temporally distinct stages from the flagger's perspective: (1) before, (2) during, and (3) after flagging. Within each of these stages, the lack of transparency in flagging creates substantial gaps in users' mental models about how platforms evaluate flagged content. Given this lack of transparency, users develop strategies to leverage their interpretation of how flags work, 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. However, flagging is not without its costs. 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 provide empirical data contributing to Crawford and Gillespie's theoretical conceptualization of flagging [15]. Our findings highlight the importance of flags on social media as a vital way for users to voice concerns within an opaque content moderation system. We also explore views of flags as a right and obligation, surfacing tensions between user motivations to flag and platforms' intentions for offering flags. We provide insights about the labor of flagging from the viewpoint of flaggers.
(2) *Conceptualization of the three stages of flagging (before, during, and after).* We provide a conceptual understanding of flagging from end users' perspectives across these three temporal stages. We highlight the key user concerns within each stage and describe how this categorization offers a useful framework to direct focused research efforts in the future.
(3) *Design opportunities.* We provide empirically informed design recommendations for innovating the current flagging systems. Crucially, we suggest incorporating seams [22] into flagging interfaces to help users understand their operations and ensure transparency throughout the flagging process. We also discuss how innovations to flagging mechanisms can address user needs while respecting their privacy and cognitive burdens.

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 identify potentially inappropriate content, which human moderators or bots subsequently review [47]. If this review determines that the content indeed violates the platform's guidelines, the flagged user may face direct consequences, such as content removal and account suspension [35, 37].

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" [42]. These personal moderation tools include *account-based tools* that let users mute or block specific accounts [24, 41, 56] and *content-based tools* that let users hide comments containing configured keywords [40] or set up content sensitivity controls [43]. Together, these mechanisms allow users to express dissatisfaction about norm-violating content (e.g., hate speech, violence, spam) [4, 48, 76] or behaviors (e.g., harassment, bullying, scamming) [78].

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* [83, 84], which let moderated users report their dissatisfaction with content moderation decisions (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) [45] or other social signals (e.g., unfollowing accounts) [74].

Users with special roles (e.g., community moderators) usually have access to a wider range of moderation tools [32, 80]. 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 [36, 90]. Similarly, Twitch streamers and YouTube creators rely on third-party applications for harmful content detection [11, 40]. 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, 12, 36] 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 asserted the importance of examining user perspectives of content moderation to inform platform practices and policymaking. Scholars argue that these internet governance regimes are defining users' roles and shaping our political and public discourse, so it is imperative to incorporate users' perspectives in their design [49, 66]. Our research on users' perceptions of flagging mechanisms highlights existing user needs to inform changes to how platforms design and administer these systems.

The discussion of user perspectives includes questions about who owns the responsibility for platform governance. Gillespie advocated for a collective civic responsibility to govern platforms [26]. However, it is not always clear how we may distinguish specific stakeholder roles when applying this form of collective governance. For instance, prior research shows that users attribute the responsibility for interventions and detection of inappropriate content to platforms [66] or human moderators [61] rather than themselves. However, Jhaver and Zhang [43] showed that users prefer personally configurable moderation tools to regulate hate speech, violent content, and sexually explicit content instead of platform-enacted bans of those categories. This 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 has highlighted substantial concerns about the burdens of physical and emotional labor that moderation places on regular end-users [42], volunteer moderators [20, 36, 51, 72, 89], and commercial moderators [70, 79], and the tensions of how users' volunteer efforts factor into platforms' monetary gains [52, 89]. While the emotional burden of moderation on volunteer and hired human moderators has been examined in detail [20, 70], we find that scant prior literature has observed the additional burden on regular users of reporting content and thus aim to explore this aspect in our study.

As an extension of discussions regarding the responsibilities to flag harmful content and the burden it places on users, it is known that people of marginalized identities are more likely to experience harm on online platforms [3, 21]. For example, Black and female-identifying users were shown to disproportionately experience an array of harms from unwanted behaviors on platforms (e.g., hate speech, doxxing) [60]. To address these challenges, researchers have prioritized the content moderation needs and perspectives of vulnerable groups [3, 31, 53, 82]. For example, Blackwell et al. examined HeartMob users' experiences on a platform built for those most affected by severe online abuse [7]. They argue that platforms' design and moderation must integrate vulnerable users' unique needs to "fully" address online harassment. This is also the rationale for our emphasis on examining the diverse experiences of users (including those from marginalized groups) with flagging mechanisms. We explore participants' concerns about flagging, focusing

on fears of retaliation, concerns about privacy, disproportionate labor required of minoritized users, and misuse of flagging mechanisms.

When examining the design of moderation systems, it is also important to consider end-users' preferences for control and transparency. Currently, most social media platforms do not offer adequate visibility into the rules and the decision-making processes behind their moderation decisions [81]. Sometimes, transparency challenges are a result of issues with existing moderation interfaces that, for example, do not offer visibility into what decisions are made by automated tools versus human moderators [44, 62]. Some researchers have examined how different configurations of moderation tool design shape the control that users perceive over their social media feeds [40, 42]. 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 [35, 38, 46] or generally increased trust [9]. 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 [18]. 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" [16] about moderation decisions when platforms fail to provide adequate or any explanations for those decisions [82]. 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) [58, 87]. This has raised concerns about the use of flagging to silence [3] or otherwise censor users [64]. 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 utilized flags to report what they defined as toxic content but also flagged teammates to whom they attributed game losses [47]. Zhao and Chen [91] reported 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 documented that users differ in what content they interpret as harmful enough to flag [39, 54], and platforms have differing guidelines on what is deemed inappropriate and thus flaggable [63]. 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 behaviors [76] that include but are not limited to trolling [25, 65], invasion of privacy [1, 8, 14, 73], public shaming [28, 71, 84], and interpersonal harm [67, 85].

Wang et al. [85] 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 [67] 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 Twitter, 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 lead Institution's IRB[1] approved this study for recruitment and data collection. We included an option to attach a screenshot of a post the user recently flagged. We also included 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 [21, 41]. 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 deeply understand 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 [31, 42].

Table 1 presents the demographic details of our participants. Although our recruitment spanned several countries, we note that all members of our research team are based in the United States. All participants indicated that they had flagged at least one social media post within the past month of their interview date. We used this experience of flagging to ground our interview discussions and let participants provide more specific feedback to our queries about their understanding of the purpose and operations of flags, their motivations for flagging, and their concerns with flagging.

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. Doing so allowed our participants to speak about their experiences and concerns with flagging, which included issues with the flagging interfaces, but also went beyond it.

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 participants describe at least one instance of

[1]We will reveal the University name after the peer review process completes.

#	Age	Gender	Race	Occupation	Country	Social Media Platforms Used
P1	35	Non-Binary	White	Digital Content Producer	USA	Facebook, Instagram, Twitter, Tiktok, Snapchat
P2	32	Female	Black	Fashion Designer	USA	Facebook, Instagram, Twitter
P3	53	Female	White	Professor	USA	Facebook, Twitter
P4	27	Female	White	PhD Researcher	France	Twitter, Whatsapp
P5	33	Non-Binary	Native American	Editor	USA	Facebook, Instagram, Twitter, Tumblr
P6	22	Female	Black	Waitress	Canada	Facebook
P7	24	Male	Black	Artist	USA	Facebook, Instagram
P8	27	Male	Black	Carpenter	USA	Facebook, Twitter, Snapchat
P9	22	Male	Black	Fashion Designer	USA	Facebook, Twitter
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, Twitter
P14	27	Male	Black	Uber Driver	The Netherlands	Facebook, Instagram, Twitter
P15	26	Male	Black	Sales Representative	Belgium	Facebook, Instagram, Pinterest
P16	35	Male	Black	Engineer	UK	Facebook, Instagram, Whatsapp
P17	27	Male	Black	Accountant	UK	Facebook, Twitter, Reddit
P18	25	Male	Black	Sales Agent	USA	Facebook, Twitter
P19	25	Male	Black	Manager	USA	Twitter
P20	34	Female	Black	Nurse	USA	Instagram, Twitter, TikTok
P21	26	Female	Black	Assistant	USA	Facebook, Instagram, Twitter
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, Twitter, Reddit
P25	24	Female	White	Unemployed	USA	Twitter, Instagram

Table 1. Participants' Demographic Information

flagging content and explain their understanding of how flagging mechanisms work. 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 recruiting, the three most common platforms participants discussed were Facebook, Instagram, and X. Specific details about which platforms each participant frequented can be found in Table 1.

We read and uploaded interview transcripts to Nvivo,[2] a cross-platform app for qualitative research. We applied interpretive qualitative analysis to all interview transcripts [59], reading and coding each interview soon after it was conducted. We "open-coded" [13] interviews in multiple iterative rounds, beginning on a line-by-line basis with codes sticking close to the transcript data. 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, we also engaged in regular

[2]https://lumivero.com/

discussions and memo writing, which helped us achieve deeper reflections and stay alert to emerging themes. After we finished processing twenty interviews, we reassessed our data and conducted five more interviews to flesh out better some themes, such as the "value of flagging" and "responsibility for flagging." At that point, our analysis reached theoretical saturation [23], and we concluded our data collection.

4 Findings

In this section, we present findings on users' mental models of flagging (RQ1), motivations to flag (RQ2), and cognitive and privacy concerns (RQ3). We begin by exploring how users learned about flagging and their understanding of its mechanisms. Next, we discuss key motivations for flagging, such as preventing harm, social pressure, and civic duty, as well as factors that discourage flagging. We present participants' feedback on flag design innovations, like making flags public or granting privileges to certain flag submitters. Finally, we report users' concerns about the burdens of flagging and potential privacy risks.

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 flagging usually noted three discrete steps: (1) before flagging (i.e., deciding if content should be flagged), (2) during flagging (e.g., categorizing the content and explaining why they are flagging), and (3) after flagging (i.e., 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."*

While the flagging interfaces and feedback shaped participants' understanding of the process, there was variability in their approaches to each of these steps. In this section, we detail findings about how our participants learned about the existence of the flagging mechanism and developed an understanding of how it functions. We delegate the discussion of user motivations to flag to Section 4.2.

4.1.1 Discovery of the option to flag content. We found that participants first learned about the option of flagging in social settings (e.g., when with friends) or while 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, and P15 was, in fact, encouraged to flag racist content by his history teacher as a way of taking action: *"I went to talk to her about [the racist content] and then she...explained that if someone did this to you...you could take action apart from just talking and nobody listening to you, you can actually flag."* Such anecdotes show how information about flagging may be spread via offline relationships users have.

Several other participants indicated that they learned about flagging while exploring platform options for addressing harm. These participants expressed a common sentiment that 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, more specifically, 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 make users aware of the choice they have to flag. However, only P3 revealed that she read platform policies to learn about flagging.

4.1.2 Review process for flagged content. Our participants provided notably different conceptions of how they believed platforms reviewed 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 of our participants were confident that platforms completely automated decisions to remove content (P2 and P7), others developed varying models of how they believed platforms balanced automation and human judgment. For the rest of the participants, there was a belief that there was some group of people who reviewed flagged content on the platform side. P19 illustrates his perception of what these people do, stating: *"[They] have people in place that work on the community guidelines...[they] 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 believed that moderators only "confirm" or "quality check" automated decisions for whether flagged content should be removed or not due to the sheer volume of content he expected the platform to handle. Participants also expressed a lack of clarity regarding how platforms operationalize their guidelines to make moderation decisions.

Additionally, we surfaced other complexities in participants' understanding of how platforms review flagged content. One theory participants developed was that platforms prioritized reviewing certain categories or types of posts over others. For instance, P15 stated that platforms *"prioritize some content over others...they base it off categories like severity."* On the other hand, P17 believed platforms used an AI to determine the review priority for a post by looking at the comments and other accounts interacting with it.

When encouraged to suggest solutions for enacting greater transparency about how platforms review flags, participants' responses alluded to a system similar to package tracking, where they would be notified immediately after they flag content that the flag submission was successful, informed about the current stage of the flag review, and then be updated about 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."*

4.1.3 Updates on flagged content from platforms. After participants submitted their flags, they described a range of outcomes. 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 their decisions to remove flagged content, making it unclear whether content had been removed or kept on the platform. For instance, P6, P12, and P17 assumed that the content they flagged had been removed when they noticed it no longer appeared in their feeds despite receiving no notification from the platform. 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 just disappeared, and YouTube never followed up with me on it. 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.

However, what seemed most frustrating to our participants with the outcomes of flagging was the **inaction** they perceived on the side of platforms. P1 felt that *"a lot of times when I flag things, they get ignored."* What it means for a flag to *"get ignored"*, however, varies. In some cases, participants would never receive any notifications, and no action would be taken on the flagged content. In other cases, platforms say they will get back to a participant but never follow up (P2 and P16). For instance, P16 noted that even though X (formerly Twitter) appeared to promise an update on its decision regarding the content he flagged, he never received a response, even weeks later. As a result of platform inaction, several participants took to voluntarily monitoring content they flagged.

4.1.4 Moderation actions performed alongside flagging. After flagging content, our participants often described engaging in additional measures prior to or alongside flagging, which revealed diverse and sometimes unconventional uses of flagging mechanisms. These actions were shaped by preexisting contexts, such as participants' roles, values, and perceived platform dynamics. P16 preferred to address harmful content by directly messaging the poster before resorting to flagging. He described his approach: *"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 their activist identity. After flagging content on Facebook, P15 and a friend confronted members of an organization behind the post in person, successfully eliciting an apology and removal of the content. This suggests that preexisting social roles and personal values significantly influenced participants' responses to content-based harm.

Other participants used existing platform interaction features to elicit greater community awareness of an instance of inappropriate content and to trigger attention for platforms to review their flag faster. For instance, P21 recounted an instance where he commented on a Facebook video depicting the sexual assault of a young woman, calling it inappropriate. He explained his motivation: *"The consequences of [me] speaking out are not as bad as those people tolerating the incidences [of abuse] that are happening."* Additionally, P19 'liked' offensive posts because he believed they would go viral, encouraging others to flag them and pressuring the platform to review the flag faster.

In some cases, participants developed strategies to disengage with content once they felt they had fulfilled their responsibility by flagging. For example, when the discussion became contentious in P21's example, he chose to unfollow the account, believing there was nothing more he could do. Similarly, 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. Our participants were keenly aware of how flags could be misused, but did not admit to misusing flags themselves. 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.

4.1.6 Concerns about platform motivations to provide flagging mechanisms. Dissatisfying experiences lead some participants to believe that flagging rarely leads to the removal of content and to speculate about platforms' motivations in providing users with the option to flag at all. Some participants surmised that platforms let users flag content for their own benefit. 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 users flagging content: they collect flagging data and use it as feedback to improve content recommendations or platform-wide moderation by funneling information to the correct decision-maker. P1 and P3 suspected that legal reasons and pressure from sources such as the Congress drive platforms to offer flags.

Participants expressed varying degrees of trust in whether flagging as a feature is designed to truly safeguard 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. They felt that platforms are motivated to protect and support users to keep the space they created safe (P1, P6, P12, and P18).

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

Our participants were motivated to flag content by both internal and external factors, such as ethical values and social pressure. Regardless of what influenced their decision, participants consistently emphasized that their decision to flag content was made independently, even when others drew their attention to potentially inappropriate material.

4.2.1 Moral outrage against inappropriate content. All participants flagged content on their feeds that they deemed harmful, basing their decisions on personal ethical judgments rather than strictly adhering to platform-defined guidelines. For many, flagging reflected a clear, pre-existing belief that certain content should be removed. P10 and P19, for instance, described relying on their own values to address whether the content was *"right or wrong"* to flag. P19 explicitly described his conscience as the *"moral standard"* for these decisions, stating *"As long as it's up to my conscience that this is right or this is not right...I flag because, according to me [and] my judgment, this post is bad."*

The act of flagging served primarily 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 users' instincts in identifying harmful posts while treating platform categories as somewhat of a framework to communicate their concerns. 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. Participants also admitted to flagging content when prompted by others. They described getting invited to flag via seeing others' comments, through direct messages (DMs), or hearing from friends through other channels of communication. 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?"*

Similarly, P6, P15, and P16 were contacted directly by people they knew who asked them to flag something. P12 noted that though she flagged comments because others urged her to, she did so only if she agreed with others' rationale. P19, however, 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 to social pressure in their flagging behaviors. Some participants were motivated to flag content by others who shared their identity characteristics. For example, P5 said: *"I have been encouraged particularly by other natives and other queer people and trans people to go ahead and report something."* Participants who were prompted to look into harmful content and flag it by others agreed to examine the content, but most emphasized that the final decision to flag the content was theirs alone.

4.2.3 Collective duty of upholding community norms. While all participants believed platforms are responsible for proactively detecting and regulating offensive content, most also thought individual users have a responsibility to flag 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)

P9 recognized a norm of generalized reciprocity [88]—they flagged content to protect others with the expectation that they would receive the same protection. This sense of responsibility was often guided by values of protecting loved ones. For example, P14 and P20 prioritized shielding their friends and family from offensive or burdensome material. As P20 explained, *"It's coming to protection. I'm protecting my kids, those people that I know."* This sense of responsibility extended to their broader community, as participants flagged content they believed could cause harm, such as racism or explicit material, with a focus on safeguarding vulnerable groups, particularly children.

4.2.4 Flagging as an avenue for user expression within content moderation systems. Participants valued flagging as a mechanism to offer their feedback. In fact, several expressed that flagging was a "right" they had as users on social media platforms. 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. A 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."* (P3)

4.2.5 Inadequate platform responses to users' flagging efforts. As mentioned in sec. 4.1.3, participants often felt that platforms' responses to their prior flags did not adequately compensate for the data or effort they provided while flagging. Such experiences demotivated them from flagging in the future. Without any feedback from platforms, participants created their own folk theories about the effects of flagging. For example, P4 suspected that Twitter automatically blocks accounts of flagged content, preventing users from seeing any content from that account again in their feed. Therefore, P4 felt blindsided when unable to see the results of her flagging efforts: *"I don't know if it's working well because I actually don't know if it's working. I have no idea what happened to that account...I have no means to actually see the status of my reporting."* In some cases, participants' interpretations of what content is flaggable were translated as what content they were confident the platform would remove. P2, P3, and P9 developed a more grim, fatalistic take on flagging, expecting platforms to ignore most flags. Together, these findings make it evident that prompt and clear feedback is essential for users to feel heard and be motivated to continue flagging efforts.

4.2.6 Strategies to flag posts versus accounts. Participants who flagged accounts exercised caution as they viewed flagging accounts as a more serious action than flagging a post because the consequences could be more severe with potential account suspension or ban (P16, P19, and P20). 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. For example, 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 attempted to infer the intentions of accounts posting seemingly inappropriate content. For instance, P19 noted that if he felt the intention of an account was good (e.g., promoting community awareness of sexism toward female athletes by posting examples of sexist language), he would not flag it. However, in contrast, he would immediately flag an account if he felt it had bad intentions.

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

In this section, we discuss participants' concerns about the cognitive load and privacy risks associated with flagging. Cognitive load concerns were particularly prominent among marginalized users who felt disproportionately exposed to harmful content. We also explored participants' views on how platforms could ease this burden by prioritizing certain flagged content for review. Finally, we examine privacy worries, including fears of retaliation from flagged users.

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 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. Transparency could help reduce the uncertainty users feel 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 asserted that users should be prioritized based on their reporting history. For instance, if users abuse flagging privileges, the content they flag in the future should be lower in priority for the platform to review. P19 also argued for prioritizing users' flags based on the amount of time they spend on the platform.

In contrast, some participants expressed that all flag contributors should be considered equally. They evoked the ideal of *equal rights for all* to support this position, arguing that having more followers should not qualify one to have more privileges when flagging. A few participants also felt that prioritizing specific users' flags is not effective as a flag review strategy. P20 considered the situation from a business model perspective of appealing to the widest demographic: *"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?* Our participants' responses to this question revealed disagreements on how flag reviews with different categories of rule violations should be prioritized. Some participants maintained that not all categories of flagged content should be considered equally. P1 and P9 argued for harmful content that is potentially life-threatening to be reviewed with the highest priority, while P15 thought racial discrimination should be at the top of the list.

Other participants opposed this view and instead held that all flags should be considered equally. P18 had a simple yet strong reason that flags should be considered equally: *"If something is not good for the community, it's not good for the community."* Similarly, P8 viewed flagging itself as a simple mechanism to generate candidates for the moderator queue, saying that *"flagging is just to bring attention to review it."* Some participants felt that establishing a priority order for reviewing different content categories was untenable. For example, P20 could not determine how the flagging system could prioritize handling racism versus sexism because of her firsthand experiences of being discriminated against both as a woman and a Black person.

4.3.3 Privacy for personally identifiable information. In addition to concerns about the cognitive burdens of flagging, participants also surfaced several privacy-related concerns. They expressed apprehensions about where flagged submissions go for processing and retaliation from opposing parties. Several participants did not want their personally identifiable information to be revealed when they flagged content. 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. In contrast, 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.

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 anecdote raises an important consideration. Though flagging is anonymous at the moment, in the aftermath, an array of concerns with retaliation arise that must be examined before new features are implemented, e.g., showing one's followers the posts or accounts one flagged. These concerns may be exacerbated if flagging were to become more transparent. To further explore this topic, we prompted our participants with the question of: *who should see flags?*

Given the hypothetical option of changing who can see when they flag content, several participants decided that all other users should be able to see when they flag content. In return, they also expected to see other users' flags. The prevalent sentiment for implementing this change was that the more times an inappropriate 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 alternative approaches, such as having only a subset of people be able to see the content they flag. P3 preferred that only a few of her followers, such as her close friends, could see the content she flags to avoid conflict since she is followed by people she has *"very different views from."* P19 supported having his followers see the content he flags, but wanted to be able to toggle this feature on and off in settings. P21 preferred that people see only some types of content she flags since she admitted she could be *"petty"* when flagging some content.

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 participants often found flagging outcomes disappointing, our analysis uncovered a strong user sentiment that flagging mechanisms are vital to their experience on social media platforms. Our participants described flagging as one of the core avenues for users to express dissatisfaction with content they deemed inappropriate. Besides flagging, our participants also noted using personal moderation tools, such as blocking mechanisms, to protect themselves from exposure to inappropriate content. Our analysis shows users 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 themselves [42].

Thus, flagging is not an isolated tool, but constitutes a key component of the sociotechnical assemblages users rely on to address harm online. Future research should examine in greater depth how users deploy it alongside or in place of other moderation mechanisms and actions. Such work may benefit from our three-stage conceptualization of flagging (described below), as other mechanisms present different affordances, cognitive burdens, and privacy concerns in their analogous stages.

The anonymity built into flagging mechanisms allows users to avoid retaliation by not having to confront perpetrators who post inappropriate content. In some cases where our participants attempted to avoid flagging by directly commenting on an inappropriate post or directly messaging a user (i.e., requesting the outright removal of another user's content), they were met with inappropriate comments or no response. They resorted to flagging as their only option for content removal. This extends our prior understanding of how users leverage their online social connections and content moderation tools [42, 90] to address online harms [75, 76].

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. Participants learned about flagging via word-of-mouth 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 demonstrates how flagging is actively employed by users for critical social functions such as protecting their friends or reinforcing commonly shared values about what is inappropriate. Our empirical insights explicate Crawford and Gillespie's [15] argument that flagging mechanisms involve complex sociotechnical arrangements, and especially point to additional avenues for researching flagging as a social phenomenon, e.g., how users learn about flagging from others (both online and offline), what shapes users' privacy-related needs and preferences when engaging with flags, and how platforms handle flag prioritization and coordinated reporting.

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. Many users viewed flagging as a *right* that platforms should uphold because it is the only avenue available for users to express that content should 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 [77], also leads users to view flagging as an *obligation*. This obligation becomes onerous when flagging requires extensive time and effort on the part of 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 [24, 42, 57, 86]. Thus, design efforts to uphold users' right in moderation systems, e.g., by enhancing control and transparency [33, 40, 81], must simultaneously attend to the additional obligations they might create, especially the emotional and cognitive labor they require.

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 [34, 43, 68]. While our findings highlight community care in the context of general community protection during flagging, Riedl and Newell [67]'s recent study showcases a specific instance of users acting to safeguard a celebrity. In that case, Taylor Swift fans, motivated by a strong sense of responsibility to protect her, proactively searched for and reported non-consensual pornographic deepfakes of the pop superstar on X—a stark example of users acting in ways platforms may not anticipate or fully support.

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. Platforms like YouTube[3] 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 [26, 77] 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

[3]https://blog.youtube/news-and-events/growing-our-trusted-flagger-program/

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.

5.3 Design Implications for Improving Flags

Our study revealed challenges across all three temporal stages of flagging. We offer stage-based design opportunities, related considerations, and examples of feature innovations in Table 2. We incorporate core values we surfaced through our analysis such as the need for greater transparency and user agency in our design opportunities.

#	Design Opportunities	Considerations	Examples of Feature Innovation
Before (determining flaggability)			
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."
During (classifying flagged content)			
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).
After (monitoring flag outcomes)			
6	Visualizing flag review status	This system should be integrated within the platform for 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 is 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 by the three flagging stages — before, during, and after flagging.

5.3.1 Before Flagging. 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.

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 [47]. Since users often 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. These educational efforts can also counter misuse, such as coordinated flagging against marginalized users [3, 58]. 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 [15]. Programs like YouTube's 2016 Heroes initiative prioritized content flagged by "trusted users," selected for their past flagging accuracy.[3] 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 [65]. 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).

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 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 underscore the need to approach flag transparency with careful consideration of user safety.

Some platforms have tested public-facing flag systems. For instance, on Zihu, users can flag content as "controversial" and engage in follow-up discussions [50]. 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 [55].

5.3.2 During Flagging. While prior research [3, 31] 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 [30] 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 [41]) 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 [40, 42]. Clearer descriptions through examples or other means can guide users in better identifying and reporting harmful content.

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[4], 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.

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.[5] 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 Bäumler et al. [5] that the German youth consider the time required for reporting as discouraging and desire a more streamlined process. Thus, we advise 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.

5.3.3 After Flagging. 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, such as who determines if the 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.

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, 19].

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 [9, 35, 38] and broader platform moderation practices [44]. However, most platforms either lack post-flagging dashboards entirely or rely on rudimentary interfaces, with Facebook being a

[4]https://transparency.meta.com/policies/community-standards/
[5]https://help.x.com/en/rules-and-policies/x-report-violation

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."[6] 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. 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.

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 [10, 49] and explored how existing policies, such as the EU Digital Services Act, address problematic platform behaviors like shadowbanning [17, 87] and unjust moderation standards [17]. Such legal and regulatory attention should extend to flagging reforms, ensuring that platforms are held accountable for transparent and equitable content moderation.

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.

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 limited us from generalizing our results and learning about the relative popularity of various design choices in flagging interfaces. We also acknowledge potential desirability bias in our interviews, as participants may have avoided discussing the misuse of flags to present themselves more favorably [6]. 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 mental models of flagging. Additionally, we recruited participants with previous flagging experience and included participants 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. Further, longitudinal methods may provide additional insights into how factors such as the volume of inappropriate content 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 [15]. 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]We were unable to find any documentation of what the dashboard looks like aside from a screenshot of flagging forum questions.

6 Conclusion

At a time in which user participation in social media moderation is being recognized as increasingly vital [40, 42, 84], our study serves as a starting point for exploring how users report inappropriate content and how mechanisms like flags can be designed to better facilitate these interactions. Our research examined why end-users flag, their experiences with flagging mechanisms, their mental models of the flagging pipeline, and reasons why they may stop flagging. We found that users perceive flags not as a stop-gap solution that compensates for platforms' limited moderation resources but as an essential avenue to voice their disagreements with the curated content. Our participants' willingness to contribute their time, even going as far as personally contacting 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 them. We call for further exploration of how flagging interfaces can be better designed with consideration of other available moderation mechanisms, the potential for misuse, and user needs for more transparent and effective ecosystems for addressing online harm.

References

[1] Esma Aïmeur and David Schőnfeld. 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] 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.

[5] Julian Bäumler, 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.

[6] 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.

[7] 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.

[8] Danah Boyd. 2008. Facebook's privacy trainwreck: Exposure, invasion, and social convergence. *Convergence* 14, 1 (2008), 13–20.

[9] 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.

[10] Nicole Buckley and Joseph S Schafer. 2022. 'Censorship-free'platforms: Evaluating content moderation policies and practices of alternative social media. (2022).

[11] 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.

[12] 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.

[13] Kathy Charmaz. 2006. *Constructing grounded theory: a practical guide through qualitative analysis.* doi:10.1016/j.lisr.2007.11.003 arXiv:arXiv:1011.1669v3

[14] 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.

[15] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428. doi:10.1177/1461444814543163 arXiv:https://doi.org/10.1177/1461444814543163

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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

[21] Maeve Duggan. 2017. Online Harassment 2017. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

[22] 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.

[23] Sandra L Faulkner and Stormy P Trotter. 2017. Theoretical saturation. *The International encyclopedia of communication research methods* (2017), 1–2.

[24] 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

[25] Carolin Gerlitz and Anne Helmond. 2013. The like economy: Social buttons and the data-intensive web. *New media & society* 15, 8 (2013), 1348–1365.

[26] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[27] Eric Goldman. 2021. Content moderation remedies. *Mich. Tech. L. Rev.* 28 (2021), 1.

[28] Lauren M Goldman. 2015. Trending now: The use of social media websites in public shaming punishments. *Am. Crim. L. Rev.* 52 (2015), 415.

[29] 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

[30] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.

[31] 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

[32] Taylor Hatmaker. 2021. Discord buys Sentropy, which makes AI software that fights online harassment. https://techcrunch.com/2021/07/13/discord-buys-sentropy/

[33] Christopher Hood and David Heald. 2006. *Transparency: The key to better governance?* Vol. 135. Oxford University Press for The British Academy.

[34] 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.

[35] 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

[36] 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

[37] 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.

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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).

[43] 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.

[44] Prerna 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.

[45] M Laeeq Khan. 2017. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in human behavior* 66 (2017), 236–247.

[46] Nancy S. Kim. 2009. Website Proprietorship and Online Harassment. https://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=1354466

[47] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 437, 12 pages. doi:10.1145/3411764.3445279

[48] 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).

[49] Kyle Langvardt. 2017. Regulating online content moderation. *Geo. LJ* 106 (2017), 1353.

[50] Chen Li. 2022. Flagging Controversies: The effect of flagging mechanisms on Zhihu platform. In *European Conference on Social Media*, Vol. 9. 237–246.

[51] 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.

[52] 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.

[53] 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.

[54] Zhifan Luo and Muyang Li. 2024. Participatory censorship: How online fandom community facilitates authoritarian rule. *New Media & Society* 26, 7 (2024), 4236–4254.

[55] 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.

[56] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. *arXiv preprint arXiv:1505.03359* (2015).

[57] 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).

[58] 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

[59] Sharan B Merriam. 2002. Introduction to Qualitative Research. *Qualitative research in practice: Examples for discussion and analysis* 1 (2002).

[60] 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.

[61] 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.

[62] 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.

[63] 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

[64] Christopher E Peterson et al. 2013. *User-generated censorship: manipulating the maps of social media.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[65] Whitney Phillips. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture.* Mit Press.

[66] 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.

[67] Martin Johannes Riedl and Ariel Newell. 2024. Reporting Image-Based Sexual Violence: Deepfakes,# ProtectTaylorSwift, and Platform Responsibility. *Available at SSRN 4919928* (2024).

[68] 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.

[69] 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

[70] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press.

[71] 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

[72] 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.

[73] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, 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.

[74] 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.

[75] 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

[76] 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.

[77] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (2022), 621–640.

[78] 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.

[79] 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

[80] 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.

[81] 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.

[82] 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.

[83] 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

[84] 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

[85] 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.

[86] 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.

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

[88] 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.

[89] 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.

[90] Lucas Wright. 2022. Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator. *Social Media+ Society* 8, 1 (2022), 20563051221077020.

[91] 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