Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms

YUNHEE SHIM, Rutgers University, USA
SHAGUN JHAVER, Rutgers University, USA

Flagging mechanisms on social media platforms allow users to report inappropriate posts/accounts for review by content moderators. These reports are pivotal to platforms' efforts toward regulating norm violations. This paper examines how platforms' design choices in implementing flagging mechanisms influence flaggers' perceptions of content moderation. We conducted a survey experiment asking US respondents (N=2,936) to flag inappropriate posts using one of 54 randomly assigned flagging implementations. After flagging, participants rated their fairness perceptions of the flag submission process along the dimensions of consistency, transparency, and voice (agency). We found that participants perceived greater transparency when flagging interfaces included community guidelines and greater voice when they incorporated a text box for open-ended feedback. Our qualitative analysis of open-ended responses highlights user needs for improved accessibility, educational support for reporting, and protections against false flags. We offer design recommendations for building fairer flagging systems without exacerbating the cognitive burden of submitting flags.

CCS Concepts: • **Information systems** → **Social networking sites**; **Social networks**; • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: platform governance, flagging, report

ACM Reference Format:
Yunhee Shim and Shagun Jhaver. 2026. Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms. *ACM Trans. Soc. Comput.* x, x, Article xxx (2026), 40 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Flagging constitutes the most visible means through which social media users can directly influence platforms' content moderation decisions. Platforms maintain complex sociotechnical systems composed of automated mechanisms and human reviewers [30, 41, 72] that continually review flagged posts and, when warranted, impose sanctions, such as removing, down-ranking, or shadow-banning the flagged posts [28, 32]. Given the expenses associated with content review [56, 72], platforms rely on flags to identify and regulate norm violations and thereby maintain the usability of their sites.

Flags are now ubiquitously available as a feature across all social media platforms, but their description, placement, and implementation vary widely. For example, Reddit offers a 'report' button below each post, which triggers a one-step selection of the post's rule violation among 14 categories, including "Hate," "Sharing personal information," and "Non-consensual intimate media." In contrast, each post on X (formerly Twitter) has a 'Report post' option embedded in a drop-down menu at the post's top-right corner; selecting this option prompts a multi-step information-gathering process about what is inappropriate in that post. These design choices constrain and

Authors' Contact Information: Yunhee Shim, yunhee.shim@rutgers.edu, Rutgers University, New Brunswick, NJ, USA; Shagun Jhaver, shagun.jhaver@rutgers.edu, Rutgers University, New Brunswick, NJ, USA.

ACM 2469-7826/2026/ -ARTxxx
https://doi.org/XXXXXXX.XXXXXXX

influence how flaggers can articulate their objections to inappropriate posts. Yet, little prior empirical research has examined how differences in flag implementations affect users' experiences of flag submissions. We address this gap by focusing on social media users in their capacity as *flag submitters (flaggers)*. Again, prior research has shown that the regular contributions of flaggers are indispensable for most content moderation systems [12, 25]—their distributed labor collectively helps moderators address the daunting task of regulating vast and evolving content [32]. Thus, it is vital that platforms design flags in ways that satisfy flaggers and encourage reporting of inappropriate posts.

In this paper, we examine how the design of four commonly deployed flagging components influences end-users. These components include *classification schemes*, which aid users in specifying their reason for the report through selecting a rule violation category; *community guidelines*, which direct users and moderators alike in assessing a post's compliance with platform standards; a *text box*, which gives users space to detail specifics of their objections; and *moderator type*, which offers users some insight into who is administering the flag review process. Current implementations of flagging mechanisms usually adopt one or more of these components [12] to assist flaggers with submitting their report and facilitate an efficient review of flagged items. Therefore, we examine how various combinations of these design components affect how users experience the flag submission process.

Specifically, this paper explores users' perception of *procedural fairness* in their interactions with flagging mechanisms. The procedural fairness perspective argues that the legitimacy of a decision-making system derives from public confidence in the fairness of the *processes* through which decision-makers exercise their authority [79, 85, 86]. Scholars widely view procedural fairness as a pivotal principle that should inform the design of social media governance systems [34, 40, 64, 74, 80]. Prior research on enacting procedural fairness in content moderation systems has examined how moderation outcomes are delivered to and experienced by end-users [40, 60, 74, 81]. Building upon this rich literature, we concentrate on incorporating procedural fairness in flagging mechanisms from the perspective of flaggers.

We aim to (a) investigate how the commonly deployed design components of flagging mechanisms impact users' perceptions of procedural fairness and (b) derive empirically informed design recommendations for improving procedural fairness in flag submissions. We examine three aspects of procedural fairness in this work—**consistency**, **transparency**, and **voice** [86]. We describe them in Sec. 2.2 and leverage them to motivate our research inquiry.

We deployed a mixed-methods approach from the outset of our study. In particular, we designed a controlled online experiment to evaluate the impact of different flagging implementations on users' fairness perceptions. We recruited 2,936 participants using Lucid Theorem[1] and randomly assigned each to experience one of 54 flagging scenarios simulated on Qualtrics. These scenarios were constructed by combining different implementations (or absence) of the four flagging components—**rule violation classification scheme, guidelines, a text box, and information about the moderator** (detailed in Sec. 3, where we motivate our specific research hypotheses). After reporting a post in such simulated flagging scenarios, participants assessed the procedural fairness of their flagging experience along three dimensions—consistency, transparency, and voice—using structured Likert scale items that our experiment presented. In addition, they responded to an open-ended question that asked how the flagging process could be improved to better support each of these fairness dimensions. This question was designed to capture users' underlying reasoning and perspectives that may not have been fully expressed through predefined scales.

[1]https://lucidtheorem.com

Our statistical analyses show that displaying community guidelines during the flag submission process raises users' perceptions of transparency. Additionally, we found that offering a text box where users can input their specific objections to the post they are reporting enhances their sense of having a voice. Our qualitative analysis of respondents' open-ended suggestions to enhance flags' procedural fairness highlights users' interest in receiving information regarding how the flag review process works and whether the flag reviewers are unbiased. Participants wanted flagging mechanisms to support greater expressivity, provide timely notifications of flag outcomes, and prevent the abuse of flags by bad actors wishing to take down otherwise appropriate content.

Our findings underscore the need for flagging systems to accommodate users' nuanced objections about why flagged posts deserve regulation. To achieve this, we recommend that flag implementations (1) integrate a text box for users to input their detailed perspectives to the platform, (2) enhance flags' vocabulary of complaints [12] by letting users highlight norm-violating portions of flagged posts and rate the violation severity, and (3) incorporate information and visualization systems that let flaggers track the review status of submitted flags. Additionally, platforms should (4) address flaggers' concerns regarding biased decision-making and false flagging by offering comprehensive information about posting guidelines, reviewers, and the review process.

2 Related Work

2.1 Content Moderation

Content moderation refers to the regulatory systems digital platforms deploy to promote effective communication among users while deterring exploitation of community attention [32]. It ensures norm compliance by promoting exemplary posts, restricting the visibility of inappropriate posts, and educating users about appropriate conduct [14, 25, 42]. Content moderation involves a range of *ex ante* and *ex post* measures—referring to actions taken before or after the content is published, respectively [30, 32]—to address content-based harms [44, 73] and improve the quality of available posts online. Over the past few years, all large-scale social media platforms have developed ad hoc content moderation infrastructures to enact these measures. Such investments have often been made in response to widespread critiques of moderation deficiencies from various stakeholders: lawmakers, researchers, news media, and the public at large [25].

Our research addresses ex post moderation, which involves actions such as content removal, downranking, demonetizing, attaching warning labels to posts, suspending accounts, and restricting the visibility of posts or accounts to specific users [28, 46]. Platforms implement such actions to reduce the negative influence of content such as hate speech [17, 69, 89], bullying [14, 24], self-harm [12, 22], violence [82], and misinformation [75, 88]. We situate flagging as a crucial mechanism within the ex post moderation stage because it identifies posts that may require moderation intervention.

While moderation decisions are made on the platform side, flagging serves as a mechanism for users to participate in the governance process [9]. *Appeal mechanisms* [86, 87], which let users express their dissatisfaction with content moderation decisions and request that they be reversed, also allow users to communicate with platform operators. However, appeal mechanisms only concern moderated users and their own sanctioned posts, leaving those flagging other users' posts with few or no options to appeal platforms' decisions.

Another bottom-up moderation mechanism that many platforms offer is *personal moderation* [44], which lets users limit the visibility of undesirable posts on their feeds. Personal moderation includes actions like muting or blocking an account and customizing the sensitivity threshold for content on one's feed according to one's preferences [44, 45]. It exclusively affects one person's feed without influencing others' content consumption [28]. Prior research examined the design

choices involved in building personal moderation tools and recommended improvements in defining interface elements, incorporating cultural context, offering greater granularity, and leveraging example content [44]. We add to this research by evaluating the design space of the flagging mechanism, another crucial user-driven moderation measure that platforms offer.

2.1.1. Flagging as a Content Moderation Tool

As platforms grow, they face the challenges of scale when enacting content review of all submitted posts [25]. Currently, flagging is employed by many social media platforms to achieve greater efficiency in their content review processes [12, 36, 75, 84]. For instance, platforms can prioritize the review of posts flagged for containing inappropriate content such as misinformation or hate speech [12, 52] and ensure immediate user safety through sanctions such as removing the flagged posts or limiting their visibility [66, 92]. Though we focus on social media platforms in this research, other digital platforms, such as online marketplaces (e.g., Amazon.com), financial apps (e.g., Venmo), and sharing economy services (e.g., Uber, Airbnb) also deploy flagging as a user-facing tool.

When we consider the impact of flagging mechanisms on end-users, three categories of users stand out: (1) *flagged users*, whose posts are (either justifiably or unjustifiably) reported by another account for infringing on platform policies, (2) *flaggers*, who request content review of selected posts or accounts, and (3) *silent bystanders*, who witness policy violations but do not flag them. Our study focuses on flaggers, i.e., users who engage with the flagging mechanism to report a post and initiate its review. While some flaggers might report content that violates platform guidelines (which is how platforms intend users to employ flags), others may use flags to express their social or political objections to the reported post [1, 52, 80], coordinate with others to collectively get a post sanctioned [75, 81], or pursue a personal vendetta against a poster [66]. In this article, we investigate flaggers' experiences using the flagging mechanism to report posts that violate platform guidelines.[2]

Zhang et al. [93] identified three temporally distinct stages associated with the use of flagging mechanisms—before, during, and after flagging—and highlighted the various user needs and interface affordances available in each stage. Our research focuses on the "during flagging" stage, and examines the design opportunities within that stage to improve users' flagging experiences.

Platforms would prefer that users flag only content that violates their existing guidelines to optimize the labor involved in content review. As such, platforms seek to concentrate flaggers' attention on reporting authentic violations of platform guidelines. They do this by designing flagging mechanisms in ways that emphasize their criteria for reviewing flagged content. For instance, flagging interfaces often prompt users to select from predefined categories of rule violations or present platform guidelines that may direct flaggers' attention to what platforms consider non-normative behaviors [8, 63]. On some platforms, flagging requires users to articulate their objections to the flagged post in a text box. This encourages users to reflect on the purpose of their flagging and elaborate on their perspectives regarding how the flagged post violates platform norms.

By incorporating these diverse elements, current social media sites demonstrate varied approaches to structuring their flagging mechanisms. For example, Instagram's flagging system lets users categorize rule violations and specify which rules the reported post has breached (Figure 1, left). Facebook offers a similar categorization interface and additionally prompts users to review the site's community standards when selecting a rule violation category (Figure 1, middle). YouTube,

[2]Note that our use of the term *flaggers* refers to general end-users using the platform's flagging mechanism; we are not concerned with *trusted flaggers*, who, as Wilson and Land [89] describe, are selected among end-users by platforms and contracted to perform moderation tasks because they possess the relevant linguistic facility or advanced knowledge of platform policies.

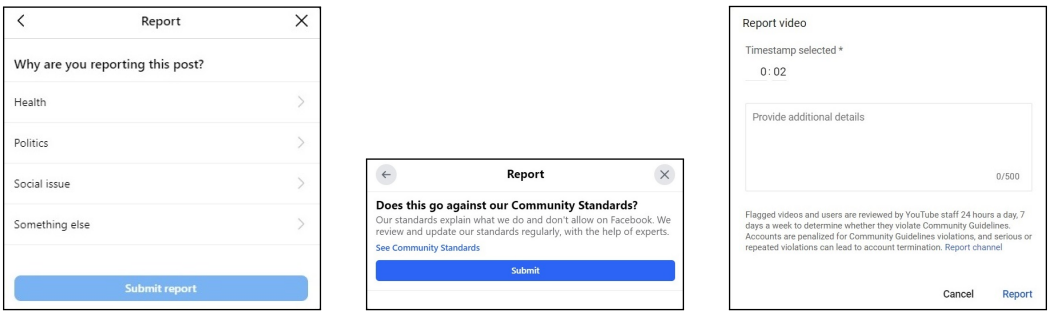


Fig. 1. When users initiate the flagging process on social media platforms, Instagram mandates that they specify the rule violation of the post (left), Facebook offers a link to its community guidelines (middle), and YouTube provides a text box to elaborate on the specific reasons for flagging content (right).

on the other hand, provides a text box, enabling users to articulate their reasons for flagging in their own words (Figure 1, right).

While these distinct components constitute the reporting mechanisms across different social media platforms, scant research attention has been paid to how each component influences flaggers' experiences. Our research integrates these diverse manifestations of commonly deployed flagging components into an experimental framework and investigates how different design choices for each component impact user perceptions. We describe three key aspects of procedural fairness—consistency, transparency, and voice—in Sec. 2.2 before turning to our hypotheses about how specific flag design components impact these aspects in Sec. 3.

2.2 Procedural Fairness in Content Moderation

Fairness is a multifaceted concept that includes procedural fairness [55, 86], outcome fairness [60, 86], and restorative fairness [51, 74], emphasizing process, outcome, and feedback, respectively. Scholars focused on incorporating fairness in decision-making systems highlight enacting unbiased processes and outcomes, avoiding unjust situations [78], and treating every individual with respect, dignity [47], and equality [59]. Content moderation scholars, in particular, have argued that achieving fairness in moderation processes entails administering equitable treatment of all users [40] through clear and consistent criteria for content review that integrates the needs of marginalized groups, such as racial minorities and LGBTQ+ individuals [33, 59].

Prior research has shown that enhancing fairness in social media platforms' decision-making systems can positively influence user attitudes [40, 59, 61] and emotions, and thereby improve the overall user experience. For instance, a higher perception of fairness regarding the moderation system enhances end-users' trust in it, which in turn lends greater legitimacy to both the decision-making process and its outcomes [70]. Notably, this holds even when moderation systems deliver unfavorable outcomes [86]. Further, moderation actions that incorporate fairness also encourage moderated users to remain active within the community [40].

Acknowledging the preceding benefits of enacting fairness in content moderation systems, researchers have explored designing systems that emphasize fairness. Previous studies have highlighted strategies such as adopting unbiased criteria for content reviews and incorporating transparency in platform protocols to enhance fairness [3, 74]. For example, Pan et al. [70] found that an unbiased moderation process could be achieved by involving an expert panel or a jury of users in the review process, which may increase users' perceptions of fairness. Several researchers have also identified communication strategies to enhance moderation fairness, such as notifying users

about content regulation via messages or emails [60] and providing comprehensive information about content removal reasons and moderator type [42].

Though this literature provides valuable insights into what influences users' perceptions of fairness in platform-enacted moderation decisions, a gap remains in our understanding of what shapes fairness perceptions of user-driven content moderation mechanisms, especially flagging systems. Therefore, we explore how flaggers' perceptions of procedural fairness may vary based on the design components of the flagging mechanism.

We focus on three key attributes to evaluate procedural fairness—*consistency*, *transparency*, and *voice*—drawing from prior research by Lee et al. [55] and Vaccaro et al. [86]. Lee et al. introduced the fairness framework, including transparency and voice as key components, for assessing algorithmic fairness [55]. Within this framework, transparency consists of standards clarity, standards validity, information representativeness, and outcome explanations, whereas voice comprises users having more control over the decision outcomes and the processes that lead to those outcomes. Vaccaro et al. developed a scale to measure fairness by drawing from both consequentialist and non-consequentialist approaches to justice, which differ in whether or not the rightness of an action should be judged based on its consequences alone [86]. They defined procedural justice as maintaining an equitable distribution of rewards or punishments and tied it to transparency, which ensures that the decision subjects understand the process, and voice, which allows individuals to express their opinions and arguments.

Inspired by these studies, we conceptualize procedural fairness in flagging systems as the adoption of clear, equitable, and consistent standard for content review, enacting transparency in delivering procedural information, and ensuring robust user participation through input mechanisms while flagging. This non-consequentialist approach to justice reflects our interest in understanding how users perceive the fairness of flag submission procedures regardless of the final decision outcomes. As such, we refrain from including aspects related to outcome explanation and outcome control in our assessment of fairness perceptions.

Below, we discuss the treatment of consistency, transparency, and voice in prior content moderation research and explain its relevance to flagging mechanisms to motivate our research questions. While the characteristics of procedural fairness are widely debated and scholars across fields such as political theory, psychology, and law have emphasized different criteria of procedural fairness [13, 48], we respond to prior calls to situate fairness within specific social contexts and acknowledge its plurality [59].

Consistency refers to the level of uniformity of the content moderation process, i.e., how content moderation policies are enacted to make moderation decisions regardless of the variable posting contexts [48, 86]. Prior research shows that Black, Indigenous, and people of color (BIPOC), LBGTQIA+, disabled, plus-sized individuals, and sex workers experience more frequent and more severe moderation consequences than other users [1, 33, 58, 59] despite being subjected to the same policies. Such inequitable moderation experiences have contributed to user perceptions that moderation systems discriminate against certain user groups based on their identity characteristics [83], further fueling the criticism that moderation has become a form of censorship [59].

Our study also addresses disproportionate moderation experiences, but we focus on a different stakeholder—*users who flag content*. We posit that users' *belief in* the platform's ability to enact consistent flag reviews, regardless of the flagger's identity, is crucial. Thus, prioritizing *cross-flagger*

review consistency[3] would mitigate disproportionality in the review process and promote a perception of moderation fairness among users. Importantly, platforms should not just enact cross-flagger consistency in their flag review processes but also satisfy end-users that such consistency exists. We examine factors that shape consistency perceptions by exploring the following research question:

RQ 1: How do design choices in platforms' flagging components affect user perceptions of cross-flagger consistency?

Note that while we focus on cross-flagger consistency, many other aspects of flagging consistency warrant attention but are beyond the scope of this study. For example, consistency regarding flag review outcomes for flagged users with different identity characteristics is important, but we chose not to examine it here because we are not concerned with flagged users in this study.

Transparency enables individuals to see how decisions are made and ensures social accountability [5, 6]. In the context of content moderation, transparency has been conceptualized as the communicative steps that platforms may take to better explain the deeper complexities of moderation processes, policies, and outcomes to the many stakeholders implicated in them [81]. Content moderation scholars have raised concerns about the lack of transparency in moderation systems, citing insufficient details provided to sanctioned users about what rules they have violated and whether their content was removed by an algorithm or a human moderator [16]. Researchers have proposed various ways to enhance transparency in content moderation systems, including providing more details about decision-making processes [40, 42, 48], describing how rule violations are detected [81], and clarifying how sanctions are determined for rule-violating posts [7].

Given that flagging is a key mechanism that drives content moderation systems, ensuring transparency about it is crucial for maintaining overall system fairness, especially since it can serve as a basis for explaining flagging outcomes. Our study focuses on enacting transparency during the flag submission process, specifically the level of detail available or implied *throughout the flag submission process* about flag reviews. That is, we conceptualize "transparency" as platforms using flagging components to inform flaggers how flag reviews are processed. We pursue the following research question:

RQ 2: How do design choices in platforms' flagging components affect user perceptions of transparency in the flagging process?

Voice refers to the extent to which users' input is accommodated in the decision-making process [48, 86]. It has been identified as crucial to shaping individuals' fairness perceptions [19]. In the context of content moderation research, Ma and Kou [59] found that YouTube creators associate the fairness of their moderation experiences with having their voice involved in the decision-making process. Despite its crucial role, limited research examines the perception of voice/agency that users experience when interacting with flagging mechanisms. We conceptualize "voice" in the context of this paper as the users' perceived degree of involvement in the flagging process, especially regarding how well they can express their objections to an inappropriate post. We thus pursue the following research question:

RQ 3: How do design choices in platforms' flagging components affect user perceptions of voice in the flagging process?

[3]Our use of this term refers to platforms' consistency in the flag review process, irrespective of the flaggers' characteristics. This should not be confused with the degree of agreement among different flaggers when flagging any content.

Table 1 presents these three attributes of procedural fairness, summarizes how previous literature has addressed each in the context of content moderation, and shows how our study conceptualizes them within the flagging mechanism.

Table 1. Procedural Fairness Attributes. For each attribute, we present its definition from prior content moderation research and our conceptualization of that attribute in the context of flagging design.

Attributes	Definition from Previous Research	Our Conceptualization for Flagging
Consistency	Platform enforces content moderation policies uniformly, regardless of the specific post context [1, 33, 48, 58, 59].	Flagging mechanism applies the same rules and standards uniformly, in accordance with the platform's values and norms, regardless of flaggers' identity characteristics.
Transparency	Platform provides users with information about moderation process and reasoning behind decision-making outcomes [40, 48, 60, 81, 83].	Flagging mechanism presents relevant information about the flagging process and flag reviews.
Voice	Platforms adopt measures to integrate users' opinion into the decision-making process [48, 59, 86].	Flagging mechanism allows users to thoroughly express their objections to the flagged posts.

Besides the four flagging components indicated above, we aimed to more deeply understand the changes that end-users seek in the broader design and policy choices associated with flag implementations. Leveraging open-ended responses in our experiment to achieve this goal, we ask:

RQ 4: How should flagging processes change to enhance users' fairness perceptions?

To summarize, our study investigates flaggers' perceptions of procedural fairness — specifically, consistency, transparency, and voice — within flagging mechanisms designed with diverse components. We situate this work in conversation with other Human-Computer Interaction (HCI) studies [2, 44, 50, 60, 70, 76] that experiment with interface and policy designs to improve platform governance outcomes.

3 Components of Flagging Design

Our study examines four components of flagging mechanisms commonly adopted by social media platforms. Drawing on the characteristics of each component and insights from previous literature, we formulate hypotheses about how these factors contribute to different fairness attributes in this section. We also describe how we operationalized each component through our survey questions.

3.1 Flag Classification Levels

As users engage in the flagging process, they may encounter a classification scheme with primary menus and submenus (see Figure 2) that specify the range of rule violations the flagged post may have, such as hate speech and spam. Making selections from predefined rule violations lets users categorize their concerns with the flagged post [12]. From platforms' perspective, imposing this precisely designed classification scheme on users is crucial to triage flagged posts, validate whether the selected rule violation occurred, and highlight that the content of flagged posts drives flag review decisions (instead of other factors, such as flaggers' identity). However, it is not yet clear how the design of classification schemes affects flaggers' perceptions of fairness.

This study explores how adjusting the granularity of rule violation selection within the classification scheme impacts users' perceptions of fairness in the flag review process. A more granular scheme

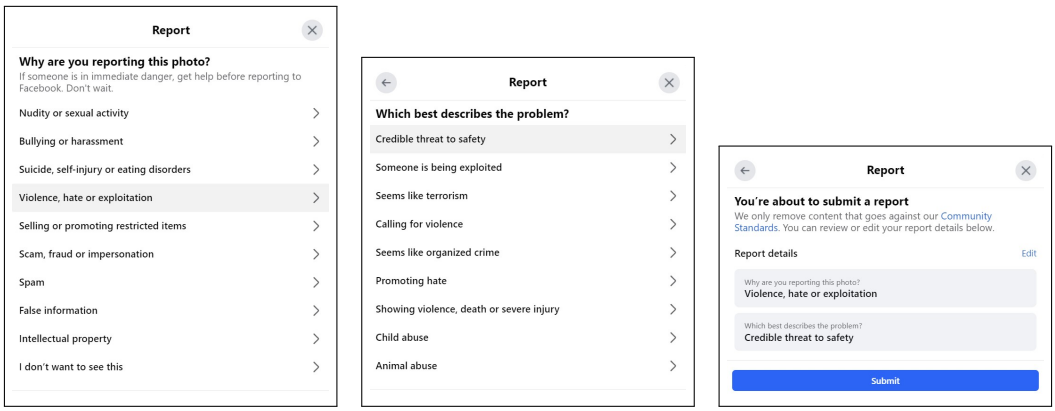


Fig. 2. This example illustrates Facebook's flag classification scheme after users initiate the flag submission process. In this example, the user categorizes her reason for flagging as 'Violence, hate or exploitation' from the main menu; this selection displays a submenu from which the user selects the subcategory 'Credible threat to safety.' Finally, the system displays these selections to the user, along with the submit button that prompts completing the report.

would let users select not only one of the few broad rule categories (e.g., hate speech, misinformation) but also the precise subcategory within each category (e.g., race-based hate speech, health misinformation).

Since the rule violation scheme is derived from content moderation guidelines, we expect that if the flagging mechanism offers users more detailed categories for flagging, they will perceive that the system prioritizes its formalized moderation criteria rather than other factors, such as flagger identity. Therefore, we hypothesize:

> *H1-1: More granular designs of flag classification schemes in a flagging mechanism will increase users' perceptions of consistency in the flag review process.*[4]

Transparency is linked to the amount of information about the process [55, 71]. Thus, furnishing a more detailed rule violation classification scheme may enhance the transparency of the flag review process. Therefore, we hypothesize:

> *H1-2: More granular designs of flag classification schemes in a flagging mechanism will increase users' perceptions of transparency in the flag review process.*

Given that voice is closely tied to providing opportunities for users' perspectives to be expressed within the process [47, 48], incorporating more detailed levels of rule violation classification scheme may lead users to perceive that the platform has a genuine interest in understanding their concerns, which may result in an elevated perception of having a voice in the process. We therefore hypothesize:

> *H1-3: More granular designs of flag classification schemes in a flagging mechanism will increase users' perceptions of voice in the flag review process.*

Operationalization

In our survey, we organized rule violations into nine distinct primary categories, each with an

[4]With this hypothesis, we also mean to include the assertion that providing even the least granular classification scheme will increase users' consistency perceptions when compared to not providing *any* classification scheme. We have similarly merged other assertions throughout Sec. 3.1 and 3.2 to achieve conciseness.

Table 2. A taxonomy of rule violation categories developed by referencing flag interfaces across multiple platforms. We used this taxonomy to operationalize flag classification choices in our survey. Each category below also included an additional submenu option of 'Something else.'

Primary category	Subdivision category
Child safety	Child exploitation
	Child neglect
	Child nudity
	Inappropriate interaction with children
False news or misinformation	Health
	Politics
	Social issue
Harassment or bullying	Me
	Someone I know
Hate speech	Race or ethnicity
	National origin
	Religious affiliation
	Social caste
	Sexual orientation
	Sex or gender identity
	Disability or disease
Impersonation	High profile impersonation
	Private impersonation
	Unauthentic behavior
Unauthorized sale	Drugs
	Weapons
	Endangered animals
	Other animals
Self-injury	Suicide
	Eating disorder
Sexual activity	Nudity or pornography
	Sexual exploitation or solicitation
	Sharing private images
Violence or incitement	Animal abuse
	Riot or terrorism
	Death or severe injury

additional submenu, as shown in Table 2. This classification scheme was developed by referencing rule violation categories in flagging mechanisms across major social media platforms, including Facebook, Reddit, X, and YouTube. By synthesizing rule violation categories across multiple platforms, we aimed to allow users to explore flagging mechanisms without focusing on the norms for any particular platform.

When designing rule violation classification schemes, we constructed three types based on the number of steps involved in rule violation category selection. Each participant interacted with one of these three scheme types:

(1) Users are not required to select a flag category.
(2) Users are required to select only a primary flag category.
(3) Users are required to select a primary flag category and a corresponding subcategory.

3.2 Posting Guidelines

Posting guidelines are crucial for designing accountable content moderation systems because they serve as the official criteria for the platform's moderation decisions [25, 62, 94]. Prior research also shows that drawing users' attention to posting guidelines contributes to pro-social community outcomes. For example, Matias [63] demonstrated through an online experiment on Reddit that

announcing posting guidelines in public discussions increases the likelihood of compliance with them and encourages newcomer participation. Further, when guidelines are easily noticeable, users are more likely to accept the platform's moderation decisions [3, 50] and perceive them as fair [40].

While this literature has established the benefits of how users' attention to posting guidelines enhances their fairness perception of moderation outcomes, little attention has been given to integrating these guidelines into flagging mechanisms and exploring its impact. We investigate how different granularity levels of posting guidelines influence users' perceptions, specifically regarding the fairness of the flagging process.

Offering clear posting guidelines in the flagging mechanism may lead users to perceive that the review process adheres to a predetermined rubric and is largely influenced by the specified guidelines and not other external factors. Additionally, we posit that more detailed guidelines, presented with specific examples of rule violations, could further strengthen the perception of a robust rule-based review process. Therefore, we hypothesize:

> *H2-1: Integrating more granular designs of posting guidelines into a flagging mechanism will increase users' perceptions of consistency in the flag review process.*

Integrating posting guidelines in the flagging mechanism is a way of disclosing content review criteria by illustrating possible flaggable posts, thereby enhancing users' understanding of the content review process [69]. We therefore hypothesize that:

> *H2-2: Integrating more granular designs of posting guidelines into a flagging mechanism will increase users' perceptions of transparency in the flag review process.*

Showing guidelines in the flagging procedure may impose a sense of the need for strict adherence to predefined rules. Since users sometimes flag in ways that may not align with platforms' notion of flaggability [52], requiring them to follow the guidelines may lead to perceptions that they cannot express their opinions freely. We thus hypothesize:

> *H2-3: Integrating more granular designs of posting guidelines into a flagging mechanism will decrease users' perception of voice in the flag review process.*

Operationalization

We developed a set of platform guidelines that explicitly describe the types of content prohibited on the platform. Table 3 shows this list and their corresponding examples. We developed these guidelines by synthesizing Facebook, Reddit, X, and YouTube guidelines. To construct them, we consulted publicly available documents such as 'community standards,' 'content policy,' 'rules and policy,' or 'community guidelines & policies,' from each platform. By identifying common guideline themes across platforms, we compiled a core list of community guidelines and incorporated specific language from different platforms to clarify them.

To construct detailed guidelines, we gathered illustrative examples for each rule violation category through two means: (1) using examples officially offered in platforms' descriptions of community guidelines and (2) conducting a content search on platforms with keywords borrowed from guidelines.

We designed three levels of posting guidelines in our survey interface, each differing in the depth of information provided:

(1) Users do not see any posting guidelines.
(2) Users see posting guidelines consisting of simple descriptions for each type of guideline violation.
(3) Users see posting guidelines consisting of simple descriptions and example posts for each type of guideline violation.

Table 3. Posting guidelines synthesized from Facebook, Reddit, X, and YouTube, and used in our survey interface. A prompt reading 'We do not allow content that' appeared above these guidelines when they were shown.

Posting Guidelines	Examples of Posts Violating the Guideline
Depicts or encourages harm against children, including maltreatment and exploitation.	-"[Sadistic video toward a child] Being strict with your child at an early age will bring you some benefits." -"Leave a child alone at home. They need to be strong by themselves."
Contains false news or inaccurate information.	- "There is no climate emergency. It's another scam. Time to wake up." - "Covid-19 vaccines can cause injury and Death. Save people from being vaccinated."
Contains bullying or threats against anyone.	- "Look at this dog of a woman! She's not even a human being — she must be some sort of mutant or animal!" - "I hate her so much. I wish she'd just get hit by a truck and die."
Demeans, defames, or promotes discrimination against individuals or groups of people.	- "A shit Muslim bigot like you would recognize history if it crawled up you cunt." - "#LGBT community is full of whores spreading AIDS link the Black Plague."
Solicits any transaction or gift of illegal/regulated goods.	- "[A picture of a firearm] Order a custom gun today—DM for purchase." - "Having cigarettes, tobacco today #Teens #studentDiscount."
Celebrates or encourages destructive behavior.	- "All my problems will disappear if I become skinnier." - "Please participate in Momo challenges [self-harm challenges] for your BEAUTY."
Contains sexually explicit images/videos.	- "[External page links]: Who wants sexual gratification? Come and enjoy!" - "Here are some [celebrity's name] wardrobe accidents & nude photo leaks. Check them out today."
Depicts or facilitates violence or aggression.	- "Here is useful information about how to hit a woman so no one knows." - "[Video showing a white nationalist punching a black BLM activist] There's no better feeling than eliminating the enemy."

3.3 Text Box

From the perspective of end-users, most content moderation processes have rigid protocols that offer limited opportunities for expression or interaction with platform administrators [44, 77]. However, some mechanisms offer flexibility e.g., a text box in a flagging mechanism that invites users to express their opinions before a flagging decision is made. We examine the utility of offering such a text box in the flagging mechanism, enabling users to explain their reasons for flagging, including how toxic posts impact them. We expect that this text box would empower flaggers to actively participate in content moderation by detailing their objections, and its unrestricted nature would let them express their opinions in diverse and personalized ways.

However, users may also perceive that incorporating text box input into the review process significantly increases the moderators' degree of freedom in evaluating flagged content, which could shift the review process away from standardized protocols. This flexibility may lead users to perceive greater uncertainty about the flag review process and how it would be shaped by the length and quality of their flag submissions. Therefore, we hypothesize:

> *H3-1: Offering a text box in a flagging mechanism will decrease perceived consistency in the flag review process.*

By elaborating the context of the flagged post and the rationale for its inappropriateness, flaggers provide moderators with a basis for post review. The availability of this text box may signal to users that moderators incorporate this input into their decisions—this offers additional insight into the review process. We thus hypothesize:

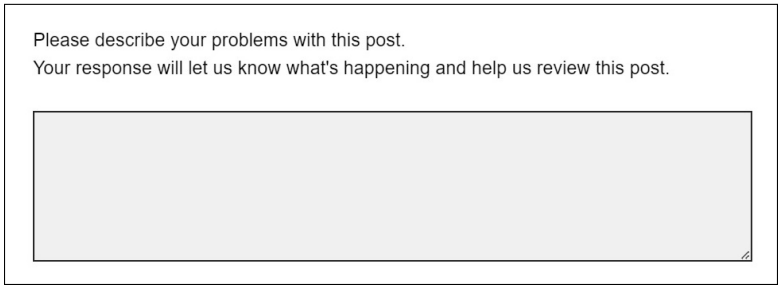


Fig. 3. Participants can express their opinions about the flagged post using the text entry feature provided by Qualtrics.

H3-2: Offering a text box in a flagging mechanism will increase users' perceptions of process transparency.

Including a text box in the flagging process lets users articulate their reasons for flagging with higher precision. This feature empowers users by granting them substantial autonomy to express their flagging intentions beyond the confines of predefined options provided by the platform's classification scheme. We thus hypothesize:

H3-3: Offering a text box in a flagging mechanism will increase users' perceptions of voice in moderation decisions.

Operationalization

We integrated a text box in our survey's flagging mechanism using the text entry feature provided by Qualtrics. This text box enabled participants to input their responses without any word limit, and it was furnished with the prompt, "*Please describe your problems with this post. Your response will let us know what's happening and help us review this post.*" Figure 3 displays how this box appeared during the survey. Each participant experienced one of the following two conditions:

(1) Users are not shown this text box.
(2) Users are given this text box to express their opinions regarding the flagged post.

3.4 Moderator Type

After users submit a flag for moderation, it undergoes assessment by either a human or automated moderator [70, 75] who evaluates whether the flagged content should be kept on the site or sanctioned. Several studies indicate a preference among users for human decision-makers over AI counterparts due to perceptions that humans can better recognize emotions [57, 90]. However, some research suggests that whether a moderator is human or AI does not significantly affect perceptions of fairness [40], transparency [29, 69] and trust [65] in moderation decisions.

Though these prior studies explore how the choice of AI versus human decision-makers shapes user perceptions of moderation outcomes, we know relatively little about how different types of moderators impact user participatory processes like flagging. We investigate how different scenarios in flagging mechanisms involving three types of moderator information (human, bot, not specified) affect users' perceptions of the fairness of flag review.

Human moderators are proficient in incorporating the context and using their domain knowledge when regulating posts, but they risk introducing potential biases in decision-making [70]. In contrast, automated moderators make decisions based on objective, predefined criteria [69], even though they may be unable to account for some relevant context. Thus, we hypothesize:

H4-1: Providing a moderator's identity as a bot compared to when it is unknown or human in a flagging mechanism will increase users' perceptions of consistency in the flag review process.

Transparency in content moderation fosters a meaningful understanding of the moderation processes and improves accountability [81]. Lack of information about the moderator type may lead users to speculate about who reviews their flagged posts, thereby reducing transparency about the process [66, 80]. Therefore, we hypothesize:

H4-2: Providing a moderator's identity as a human or bot compared to when it is unknown in a flagging mechanism will increase users' perceptions of transparency in the flag review process.

Prior studies have shown that users perceive decisions made by human moderators, such as expert juries, to be more legitimate than those made by AI [70]. Additionally, users prefer expressing their opinions to human moderators during the appeal process since they believe that humans are more likely to examine their opinions with empathy and compassion [91]. We thus hypothesize:

H4-3: Providing a moderator's identity as a human compared to when it is unknown or a bot in a flagging mechanism will increase users' perceptions of voice in the flag review process.

Operationalization

Following the other flagging prompts, we included a thank you message to users for their flagging action in our survey, optionally accompanied by moderator information. Specifically, each participant saw one of the following messages:

(1) No moderator information: "Thank you for reporting."
(2) Human moderator review: "Thank you for reporting. Our team of human moderators will review your flagged content."
(3) Automated moderator review: "Thank you for reporting. Our automated moderator will review your flagged content. Automation enables us to manage the hundreds of millions of reports we receive annually more effectively."

4 Methods

We begin this section by describing how we implemented 54 distinct flagging scenarios using Qualtrics to survey our participants. Next, we offer information about our survey deployment and detail our ethical considerations. Finally, we describe the methods we used to analyze the Likert scale and open-ended responses from our survey data. We preregistered this study at OSF.[5]

4.1 Survey Design

Our mixed methods study deployed a triangulation design [37] in which we collected both quantitative and qualitative data using a questionnaire that contained both open-ended and close-ended questions. Both methods were given equal weight [37] and the mixing of the two methods occurred during the interpretation of the study results (Sec. 6). This approach allowed us to compare and contrast our findings to produce well-validated conclusions and gain an in-depth understanding of how the overall design and procedural space of flag mechanisms shape flaggers' fairness attitudes.

To collect data for this study, we created an online survey using Qualtrics software and distributed it via Lucid Theorem, an academically-oriented survey platform, to collect responses from

[5]https://osf.io/982wd/?view_only=91e6792f777b44b582fd6bb480ab142c

U.S. adult internet users. The survey began with a page displaying seven examples of rule-violating posts, shown in Table 4. ***Warning:*** The example posts in Table 4 may be disturbing to some readers.

Table 4. Rule-violating Posts that Survey Participants Were Asked to Choose From to Initiate the Flagging Mechanism.

No.	Post
1	Mexicans come from an uncivilized, backward society. They are filthy criminals, molesting innocent American women and menacing entire neighborhoods. For the sake of our safety, they should all be beaten up and rot in jail forever.
2	@Sean11 I hate all you F***er and your mother all of you should go die in a hole!
3	Drinking bleach has been scientifically proven to prevent COVID-19. You should drink a 250ml glass of bleach with every meal. We will get through this together!
4	Today is a great day! I ate 723 calories today and have been exercising since 10 a.m. It's better to die than to never be skinny. Maintaining 40 pounds is a mandate for an adult woman. #Anorexia #annorexya #anoreskya
5	Download the software for credit card number generator today - [website URL]
6	Omg just got tons of Bucks from here! - [xyz phishing site].com.
7	I have a masturbation video of @Janny12. **Cute Yellow Chink Asian! ** Selling for $100 for just 5 minutes! If anyone wants 37 minutes of her video, send me a private message.

We borrowed and modified these rule-violating examples from previous research [18, 29, 53, 68] as well as posts we encountered during our regular use of social media. These posts reflected the diversity of norm violations users might encounter in their everyday use of social media, including instances of hate speech, bullying, self-injury, and misinformation. We constructed three of these posts in ways that pose classification challenges due to either the absence of a clear rule violation category or multiple possible selections within the rule violation classification scheme. For instance, post #7 in Table 4 is an example of race-based hate speech that is also sexually explicit. This approach aimed to encourage participants to consciously explore the flagging process and invest more effort into elaborating upon their intentions.

The survey instructed participants to read all seven posts and choose one to report. At this point, we also described the concept of flagging on social media, explained that our survey questions would simulate the steps involved in flagging a post, and requested participants to follow these steps to submit a flag in ways similar to how they would flag their selected post on a social media platform. After selecting a post, participants were randomly assigned to one of the 54 flagging scenarios simulated via Qualtrics questions (i.e., we did not direct participants to an external site) as described below. This selected post was displayed at the top of the screen to guide participants throughout the flagging procedure.

The flag submission choices shown to participants followed a between-subjects factorial design created by combining different implementation choices for the four flagging components. It comprised 3 (Rule violation classification schemes) $\times$ 3 (Granularity of posting guidelines) $\times$ 2 (Availability of a text box) $\times$ 3 (Types of moderators) different conditions. Depending on their assigned scenario, participants experienced variations in the design of the rule violation classification scheme, the level of detail regarding platform guidelines, the presence or absence of an open-ended text box, and information about the moderator (human, bot, or left unspecified).

4.1.1. Flagging Scenarios

Section 3 details how we operationalized the four components of flagging designs within the Qualtrics software. In summary, Figure 4 illustrates the flagging components and an example of

flagging design flow a participant could have encountered in the survey. The left column enumerates each component and its different levels (i.e., conditions). The right column shows one example of the flagging scenario a participant might have encountered. Once a participant selected one of the seven inappropriate-post examples listed above, the flagging process (scenario) started. In the scenario illustrated in Figure 4, a participant would have been required to choose a rule violation from the primary classification scheme for the post they selected. Then, the participant would have been offered a text box to describe the rule violation in detail. The final step would have been a message of gratitude for flagging, informing them that an auto-moderator will review the post they flagged.

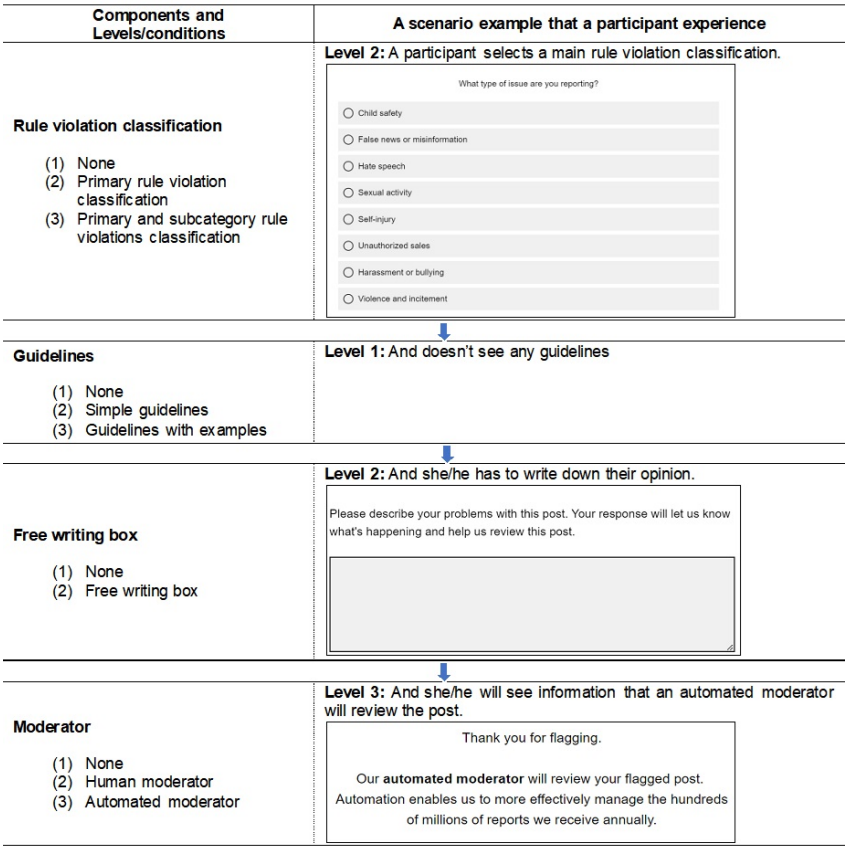


Fig. 4. The four components, the different levels of each component, and an example scenario of the flagging mechanism experienced by participants. The left column outlines each element of the flagging mechanism along with its corresponding levels, while the right column exemplifies a scenario comprising a randomly selected combination of these levels.

4.1.2. Dependent Variables: Procedural Fairness

In each scenario, after participants completed the flagging, they were prompted to evaluate their perceptions of procedural fairness regarding the flag mechanism they encountered. The survey directed participants to respond to three questions assessing procedural fairness that we developed for this study, as outlined below:

(1) "This flagging mechanism will review each flagged post consistently regardless of the flagger's personal characteristics, including gender, age, and account history."
(2) "This flagging mechanism offers sufficient visibility into the important elements of the flag review process."
(3) "This flagging mechanism allows me to fully express my objections with the flagged post."

We used responses to these questions to operationalize participants' perceptions of consistency, transparency, and voice, respectively. Participants indicated their responses to each of these questions on a 7-point Likert scale that ranged from *Strongly Disagree* to *Strongly Agree*.

Additionally, to answer RQ 4, the survey featured three open-ended questions asking participants to provide suggestions on how the flagging process could be further enhanced. These questions were:

(1) "What changes would you suggest to this flagging mechanism process to increase your trust that the flag review is consistent regardless of who flagged the post?"
(2) "What changes would you suggest to improve the transparency of this flagging mechanism process?"
(3) "What changes would you suggest so that users can more fully express their objections about the concerned posts?"

4.1.3. Usability

As detailed above, our experimental design enacted 54 distinct flagging scenarios, each comprising a combination of varying levels of components, resulting in differing workloads for each scenario. According to the Technology Acceptance Model (TAM), which examines user interaction with systems, the cognitive effort expended during system use significantly influences subsequent system adoption [15]. In accordance with this model, we expect that the diverse scenarios in our survey could impose varying cognitive burdens on the users and influence their future use of flags. Thus, in each scenario, we included two questions to assess usability as part of the survey: one regarding the effort required by users to use the flagging mechanism and the other regarding participants' intentions to use this mechanism in the future. Participants were asked to rate the following two questions using a 7-point Likert scale:

(1) "How demanding was it for you to use this flagging mechanism to report the post you selected?" (rated on the scale of 'Very demanding' to 'Very undemanding')
(2) "How likely are you to use this flagging mechanism to report any inappropriate post on a social media platform you use?" (rated on the scale of 'Very likely' to 'Very unlikely'.)

4.2 Deployment of the Survey

Our study was considered exempt from review by Rutgers University's IRB on 16 November, 2023. Before launching the survey, we conducted a pilot survey from 1 May to 8 May, 2024 to gather feedback on the survey's wording and organization. Based on this feedback, we revised the survey questionnaire by adjusting certain words and adding signposts. This piloting also gave us confidence that the participants, including those without any prior flagging experience, could immerse themselves in the flagging task during the course of the survey. After this, we rolled out our main survey, which targeted U.S. adult internet users aged 18 and above. Participants were recruited through Lucid Theorem (https://lucidtheorem.com), a survey company providing access to nationally representative samples, from 16 May to 18 May, 2024. Participants who completed the survey received $1.50 in compensation via the Lucid system. In total, we collected responses from 3,650 participants.

We excluded respondents who opted out during the survey (N = 496).[6] We also conducted several pre-processing actions to ascertain our data quality, e.g., we filtered out responses where participants spent less than 1 minute or more than 50 minutes on the survey (N = 40), or exhibited straight-lining behavior by selecting the same response for all questions (N = 178). Following these pre-processing steps, we retained 2,936 survey responses, which we used for our subsequent analyses. Appendix A describes the demographic distribution of our survey sample. Note that while our use of Lucid Theorem let us recruit demographically diverse participants, our research concerns with *process inference* rather than *population inference* [35, Ch. 3], i.e., we are more interested in illuminating the processes producing the effects observed rather than guaranteeing the population representativeness of our data providers.

On average, the participants took 393.2 seconds to complete the survey. As mentioned above, the survey began with participants selecting a post to flag from the list of rule-violating posts presented in Table 4. Table 9 (Appendix B) shows the distribution of post choices made by survey participants and how average fairness perceptions vary across these choices — we did not find any noticeable trends in fairness perceptions based on participants' post choices.

4.2.1. Ethical Considerations

Our survey contained several examples of social media posts that breach platform guidelines and are regarded as inappropriate. These cases were included to direct participants' attention to the rationale behind reporting and to provide a realistic context for navigating the flag submission procedures. When obtaining consent to participate in this study, we informed participants of the potential risk of encountering inappropriate posts before they began the survey. Participants who preferred not to view these posts could choose not to participate and opt out of the survey.

We included only textual content in our examples, refraining from using visually disturbing material to mitigate the risk of participants experiencing psychological harm. To safeguard the mental health of our participants, we provided information about mental health resources, including the contact information of organizations such as the National Institute of Mental Health and Mental Health America, for participants to use if they needed help.

4.3 Data Analysis

4.3.1. Quantitative Analysis

We used SPSS Version 29 to analyze our quantitative data. Sec. 5.1 presents the results of our quantitative analyses. To test hypotheses *H1-3*, *H2-1*, and *H2-3*, we conducted one-way Analysis of Variance (ANOVA) tests. ANOVA examines whether statistically significant differences exist between three or more groups. In cases where homogeneity of variance was violated (Levene's test resulted in a p-value below .05), we employed Welch's ANOVA test (*H1-1*, *H1-2*, and *H2-2*). For testing hypotheses *H3-1*, *H3-2*, *H3-3*, *H4-1*, *H4-2*, and *H4-3*, we conducted independent samples t-tests. This method examines whether statistically significant differences exist between two groups.

As an additional exploration, we built a General Linear Model (GLM) to examine the interaction effects among our independent variables on participants' perceived consistency, transparency, and voice. This analysis demonstrated how different combinations of component choices impact users' perceived fairness. Additionally, we evaluated how each flagging component impacted flaggers' cognitive burden. To test whether different choices of classification schemes, posting guidelines, and moderator type affect participants' cognitive load, we conducted three separate ANOVA tests,

[6]Our review of collected data shows that 253 participants opted out on the first page where we requested participation consent, 7 dropped out at the next question requesting selection of a post that participants would like to flag, 148 dropped out in the flag submission stage, and 88 dropped out after answering the flag submission questions. We found no noticeable trends in the number of dropouts across the different flag implementation scenarios.

one for each of those components. In the case of the text box, we conducted an independent samples t-test to see how the availability of the text box affects users' perceived cognitive burden.

4.3.2. *Qualitative Analysis*

The survey included open-ended questions (Sec. 4.1.2) to elicit suggestions for enhancing consistency, transparency, and voice in the flagging mechanism. After cleaning the data, we collected 1,741 valid responses for these three questions: Consistency (N = 657), Transparency (N = 691), and Voice (N = 393). Next, we performed an inductive analysis [11] on these responses using NVivo v.14.

Although the survey provided separate boxes for suggestions related to each aspect of fairness (consistency, transparency, and voice), respondents frequently combined and expressed their responses in a single long answer, highlighting the interconnected nature of these aspects. Thus, we integrated responses to these separate sections and analyzed them together to understand the participants' intentions better. In addition, we excluded the responses that merely pointed out the importance of each flagging component without any additional elaboration, e.g., a response just stating 'text box' was excluded. This step helped us surface more nuanced insights.

Next, the two coauthors independently analyzed the initial 20% of open-ended responses concerning consistency, transparency, and voice, and coded them. Some responses were so detailed and complex that we attached multiple codes to them. Subsequently, we engaged in iterative discussions to compare and refine our codes, often reflecting on emerging concepts and quoted responses, and achieving consensus on code definitions and applications through collaborative discussions. After coding the entire dataset, we refocused our analysis at the broader level of themes rather than codes. Specifically, we iteratively sorted related codes into potential themes by using mind-maps and attending to the relationships between codes and between emerging themes [4]. Through this process, we next reviewed and refined our candidate themes and arrived at five key themes we present as our findings. Sec. 5.2 presents the results of our qualitative analysis.

5 Findings

5.1 Quantitative Findings

Table 5. Mean (M) values of Perceived Consistency, Transparency, and Voice Across Different Conditions of Flagging Components (Likert Scale: Strongly Disagree = 1 to Strongly Agree = 7). Standard deviation (SD) values are in parentheses. *N* represents the number of participants placed in each condition.

Condition Group	Condition	*N*	Consistency (M, SD)	Transparency (M, SD)	Voice (M, SD)
Classification Scheme	None	968	5.53 (1.45)	5.29 (1.43)	5.23 (1.66)
	Simple	980	5.61 (1.39)	5.35 (1.32)	5.37 (1.45)
	Detailed	988	5.48 (1.51)	5.27 (1.39)	5.30 (1.52)
Guidelines Level	None	993	5.51 (1.48)	5.15 (1.45)	5.24 (1.61)
	Simple	971	5.53 (1.45)	5.33 (1.37)	5.31 (1.54)
	Detailed	972	5.59 (1.43)	5.44 (1.29)	5.35 (1.49)
Text Box Availability	Absent	1,492	5.54 (1.44)	5.27 (1.38)	4.94 (1.64)
	Present	1,444	5.55 (1.46)	5.33 (1.38)	5.67 (1.35)
Moderator Type	Not Available	988	5.56(1.46)	5.32 (1.38)	5.33 (1.54)
	Human	972	5.52(1.45)	5.27(1.41)	5.26 (1.55)
	Bot	976	5.55 (1.45)	5.33(1.34)	5.32(1.55)

Table 5 shows the mean values of perceived consistency, transparency, and voice reported by participant groups that encountered different conditions for the four flagging components. Given

that we ran three comparison tests for each flagging component (one each for consistency, transparency, and voice), we estimated the statistical significance of each result following Bonferroni correction (α < .05/3). Table 6 summarizes the results for our hypothesis.

Table 6. Summary of Research Hypotheses, Variables, and Outcomes

Hypothesis	Independent Variable	Dependent Variable	Outcome
H1-1 More granular classification schemes will increase perceived consistency.	Granularity of classification scheme	Consistency	Not supported
H1-2 More granular classification schemes will increase perceived transparency.		Transparency	Not supported
H1-3 More granular classification schemes will increase perceived voice.		Voice	Not supported
H2-1 More granular posting guidelines will increase perceived consistency.	Granularity of posting guidelines	Consistency	Not supported
H2-2 More granular posting guidelines will increase perceived transparency.		Transparency	**Partially supported**
H2-3 More granular posting guidelines will decrease perceived voice.		Voice	Not supported
H3-1 Offering a text box will decrease perceived consistency.	Text box availability	Consistency	Not supported
H3-2 Offering a text box will increase perceived transparency.		Transparency	Not supported
H3-3 Offering a text box will increase perceived voice.		Voice	**Supported**
H4-1 A bot moderator will have higher perceived consistency than an unknown or human moderator.	Moderator type	Consistency	Not supported
H4-2 A bot or human moderator will have higher perceived transparency than an unknown moderator.		Transparency	Not supported
H4-3 A human moderator will have higher perceived voice than a bot or unknown moderator.		Voice	Not supported

5.1.1. Rule Violation Classification

We tested Hypotheses 1 by examining how different implementations of rule violation classification schemes affect user perceptions through three separate ANOVA tests. First, we rejected ***H1-1*** given that contrary to our hypothesis, participants shown a simple classification scheme have a higher perceived consistency than those shown a detailed one (Table 5). Further, no significant differences were observed between the absent and simple conditions, nor between the absent and detailed conditions. We also rejected ***H1-2*** since the different classification schemes do not significantly impact perceived transparency. Finally, we rejected ***H1-3***, given that differences in classification scheme do not significantly impact perceived voice.

5.1.2. Posting Guidelines

To test Hypotheses 2, we used three ANOVA tests to examine how displaying posting guidelines with different granularity levels affect user perceptions. We found that these differences do not significantly impact participant perceptions of consistency and voice, thus rejecting ***H2-1*** and ***H2-3***. However, our Welch's ANOVA test confirmed a significant relationship between the levels of guidelines and perceived transparency ($F(2, 1952.81) = 11.36$, $p < .001$). Conducting a Games-Howell post hoc analysis, we found partial support for ***H2-2***—participants encountering simple (M = 5.33) or detailed guidelines (M = 5.44) have a significantly higher perception of transparency than those in the absent guidelines condition (M = 5.15).

5.1.3. Text Box

We conducted t-tests to examine how the availability of a text box affects user perceptions to test Hypotheses 3. Our analysis shows that the availability of open-ended text boxes that let users describe their reasons for flagging does not significantly affect their perceptions of consistency

and transparency, leading us to reject ***H3-1*** and ***H3-2***. However, we found support for ***H3-3***—participants in the text box condition (M = 5.67) report a significantly higher perceived voice compared to those without such a box (M = 4.94, $t(2859.50) = -13.32$, $p < .001$).

5.1.4. Moderator Type

To test Hypotheses 4, we conducted t-tests to examine how different moderator types affect user perceptions. Our analyses found that different moderator identities do not significantly impact perceived consistency, transparency, and voice. Therefore, we reject the hypotheses ***H4-1***, ***H4-2***, and ***H4-3***.

5.1.5. Additional Analyses: Interaction Effects and Usability

Interaction Effects. In addition to testing our proposed hypotheses, we explored the interaction effects among the four flagging components on perceived consistency, transparency, and voice. Table 10 in Appendix C shows the results of this analysis. We estimated the statistical significance of these results following Bonferroni correction ($\alpha < .05/15$) and concluded that the interaction between classification schemes and the availability of a text box significantly affected perceived voice ($F(2, 2936) = 8.29$, $p < .001$). Specifically, for participants in the condition *without* a text box, providing either a simple (*MD (mean difference)* = 0.34, $t(2882) = 3.48$, $p < .001$) or detailed (*MD* = 0.35, $t(2882) = 3.59$, $p < .001$) classification scheme elicited significantly higher perceptions of voice than providing no classification scheme (Table 11, Appendix C). All other interactions did not significantly affect any of the three aspects of procedural fairness.

Usability. We examined how variations in each component of flagging mechanisms differentially impacted the cognitive burden on our participants, thereby potentially influencing flag usability. Our analysis (Table 12, Appendix C) shows that participants in the condition with a text box ($M = 3.86$, $SD = 1.73$, $SE = .05$) experienced a significantly higher cognitive burden compared to those without such a box ($M = 3.66$, $SD = 1.80$, $SE = .05$). No other effects approached statistical significance. We also found that variations in how each flagging component is presented (or not presented) did not significantly affect participants' self-reported likelihood of using flagging mechanism (Table 13, Appendix C).

5.2 Qualitative Findings

Our inductive analysis surfaced five themes through our combining and distilling of codes that together provide a nuanced understanding of user perspectives on enhancing fairness in flagging mechanisms, thus addressing RQ 4. Table 7 summarizes these themes and their frequency. These findings offer additional insights into the implementation of flagging components we explored in our quantitative analysis, but they also offer a broader perspective on other design and policy solutions that could enhance fairness in flagging systems.

Table 7. Analysis of Responses to Open-ended Questions: Suggestions for Enhancing Fairness.

Theme	Frequency
Desired attributes of flag reviewers	138 (11%)
Needing support for greater expressivity	297 (24%)
Demanding outcome notifications with timely review	361 (29%)
Expectations regarding review procedures and statistics	298 (24%)
Preventing flagging abuse and protecting flaggers	144 (12%)

5.2.1. Desired Attributes of Flag Reviewers

Participants frequently expressed their concerns about flag reviewers' biases and detailed their preferences for either human or AI moderators and their reasons for such preferences. Some participants (N = 61) mentioned that they **prefer a human reviewer** who can understand the context and nuance of the flagged post. Further, a few (N = 9) expressed their interest in **communicating with moderators** about the flagging process and the results. This suggests that users need thorough explanations for flagging outcomes through detailed conversations. For instance, participant P312 suggested:

> *Enable direct messaging with moderators for in-depth discussions.*

However, other respondents (N = 24) expressed a **preference for AI-based reviewers**, such as a bot or an algorithmic moderator, to enhance fairness. These divergent views regarding the types of reviewers are consistent with our quantitative findings, which indicate no significant differences in perceived fairness based on the moderator type.

Many respondents offered suggestions to reduce biases in decision-making. For example, 17 participants suggested that platforms implement collaborative decision-making by involving multiple moderators for flagged posts. Some participants specified their suggestions for a certain combination or number of moderators (often three) or a mix of human and bot moderators. They believed that such cooperative decision-making in the flagging process would increase fairness and reduce biases. For example, P405 wrote:

> *Form a diverse review team comprising individuals with varying backgrounds, perspectives, and experiences. This diversity helps mitigate biases and ensures a more balanced evaluation of flagged content.*

Further, to reduce a biased flag review process, 27 participants emphasized the desired qualifications and characteristics of individual moderators. This includes moderators' expertise in the subject matter of reviewed content, their understanding of ethical standards regarding rule violations, and their political neutrality.

5.2.2. Needing Support for Greater Expressivity

Respondents often noted that the existing flagging features do not allow them to fully express their objections to the flagged posts, and they proposed many constructive suggestions for improving flagging systems to better support expressivity. Among these, the most frequently suggested feature referenced the classification scheme, such as adopting a **broader classification scheme** (N = 45) with a wider range of rule violation categories for selections and the inclusion of **multiple choices for flagging reasons** (N = 90), which would enable flaggers to select several rule violation categories within the classification scheme. For example, Participant P158 advocated for allowing the selection of multiple categories, anticipating that it would induce changes at the flag review stage:

> *Allow to select more than one answer. Sometimes flagging should be reviewed for more than one concern: this might increase the likelihood of a thorough review.*

Such responses reflect participants' desire to provide the platform with a more precise understanding of their objections with flagged posts by elaborating on the types of rule violations they encounter. Other suggestions on how to achieve this included an ability to **rate the severity of posts** (N = 15) or **highlight specific sections of posts that violate rules** (N = 14). For instance, P708 noted:

> *Even within flagged items, the severity level is not touched. Some flagged categories are more harmful than others.*

In addition to emphasizing their need for clear and detailed reporting, some participants stressed the importance of maximizing user participation in the flagging process. For instance, they desired an ability to **use alternative channels** (N = 21) for flagging inappropriate posts, such as chat, email, or phone calls. Some felt that platforms should **encourage users to flag more** (N = 19). This could be achieved by providing clear instructions, step-by-step explanations, or incentivizing flag submissions for users. Another suggestion to enhance user participation and promote inclusivity was to **simplify the flagging process** (N = 73), ensuring that it is user-friendly and accessible to a wide range of users.

> *Use easy words to describe so that everyone will understand easily. (P782)*

Further, some participants advocated for adopting measures that enable flaggers to corroborate a post's toxicity by sharing their perspectives and referencing previously flagged posts. This finding suggests that incorporating multiple perspectives, not only from flag reviewers but also among flaggers themselves, could contribute to designing a fairer flagging mechanism. As specific methods, respondents expressed interest in **discussing toxic posts in a forum** (N = 21), where they could share their thoughts on the regulation of such posts. Another approach to ensuring accurate judgment of the post's toxicity involved **reviewing similar previously flagged posts and their post-review outcomes before submitting a flag.** Participants believed that implementing such features would enhance the credibility of their objections and provide the platform with clear insights into the urgency of removing flagged content.

Besides communicating their objections to individual posts, some participants desired an ability to **provide feedback on their flagging experience** (N = 6). That is, they wanted to voice their opinions not only about inappropriate posts on the platform but also about the flag submission process and its results. P249 elaborated on how the platform must listen to users' opinions on the flagging mechanism:

> *The [flagging] mechanism should be reevaluated at least every three months in case any revisions are indicated over time.*

This shows that participants want to contribute and express their perspectives on how the flagging mechanism should operate. In line with this, some participants recommended including a **rebuttal or appeal option in the flagging process** that would allow flaggers to contest the flag outcome decisions (N = 12). P708 specifically emphasized that an appeal process for flagging outcomes is crucial for both parties—the flagger and the author of the flagged post:

> *You should be able to dispute the flag on both sides of it.*

5.2.3. Demanding Outcome Notifications with Timely Review

Many respondents (N = 284) expressed a strong preference for receiving **notifications of flagging outcomes and explanations** for those outcomes. Given that our survey questionnaire was designed to gauge perceptions of procedural fairness prior to decision-making for the flagged content, this finding highlights that participants closely associate the decisions regarding flagged posts with the overall fairness of the flagging mechanism. P312 emphasized the importance of transparency in decision-making, stating:

> *The review committee should write back to the flagger why they agreed or disagreed with the flagger's decision to flag certain content.*

Some respondents expressed a strong desire to receive a timely reply to their flags, which relates to the 'voice' aspect of procedural fairness. This suggests that flagging mechanisms should not only allow users to express their concerns during flagging but also ensure that these concerns are promptly considered and addressed in a responsive manner. This desire was especially reflected in

participant suggestions for an **expedited review process** to inform flaggers of flag outcomes (N = 31). For instance, P174 noted the importance of timely reviews, recommending that:

Ensure follow-up or decision is made within 24 hours.

Additionally, some respondents (N = 46) called for **immediate actions**, such as post removal or account suspension upon flagging. For instance, P304 suggested making flagged posts invisible until a decision is reached about their removal:

Delete the post until you confirm it is something that violates the rules.

To sum up, users perceived fairness in the flagging process as being related to the platform's immediate responsiveness and timely notification of the flagging results. We infer from these suggestions that users' perceptions of fairness *during* flagging may depend on their beliefs of how seriously the platform takes their flagging requests, how clearly the decision outcomes are explained, and the time it takes to review the requests.

5.2.4. Expectations Regarding Review Procedures and Statistics

Respondents often desired to learn more about platforms' processing of flags at the system level. In addition to the flagging results, a significant portion of participants (N = 131) shared their need for more information on the flagging process, specifically about **how the review process works**. They argued that the flagger should be informed about the review criteria the platform adopts, who conducted the review, what steps are entailed in the review process, and how long they should expect to wait until the decision-making occurs. For instance, P500 proposed:

Probably more specifics. "Your report will be reviewed with[in] X period of time."

In addition, 15 respondents mentioned that the flagging mechanism needs to provide **information about the possible outcomes of flag review**, such as the removal of flagged posts and the additional sanctions that flagged users may face. They felt such information would help end-users decide whether and how to flag inappropriate posts they encounter. Further, 137 respondents conveyed their specific needs to have regular updates on the submitted flag by **tracking changes in the flag review status**. They desired a flag tracking system that could offer information such as the confirmation of flag submission, the review steps already taken, the current review stage, and the remaining steps. They hoped that such information could be provided via emails, message notifications, or a dedicated flag tracking page on the platform. For instance, P540 mentioned:

[Platforms should] offer an incident number that the person making the flag can refer to, then allow a process to follow the incident number through completion.

Some participants suggested that platforms provide **statistics on flag submissions and flag review outcomes** (N = 15). They felt that platforms could enhance transparency about flagging by disclosing statistics related to the number of flags submitted and deleted as well as the relative frequencies of rule violation categories of submitted flags in the form of monthly or annual reports.

5.2.5. Preventing Flagging Abuse and Protecting Flaggers

In addition to the suggestions about various components of the flagging mechanism, respondents provided feedback on improving overall platform management, including the flagging mechanism. Some respondents (N = 20) emphasized the need to **prevent abuse of flagging** by malicious users. This concern was specifically raised about potential misuse, where flags might be submitted without any valid reasons for post removal. They noted that the platforms must verify each flag to ensure it originates from a human user, i.e., it is not automated, and that it is not driven by specific political agendas. P109 underscored this by stating:

Make sure the flag is correctly verified because some people just flag others for no reason.

Additionally, respondents advocated for implementing ex-ante **preventive measures** (N = 38) to curb the dissemination of toxic content, and thereby eliminate the need to flag it ex-post. Suggestions included employing advanced filtering systems that block posts containing certain keywords and displaying some sanctioned posts as examples of what constitutes norm violations for educational purposes. Some respondents also suggested that flaggers should **directly communicate with the authors of rule-violating posts** independent of the flagging mechanism, which could reduce reliance on flagging as the primary means of enforcement. P816 suggested:

> *Instead of reporting, go directly to the person and talk to them. If that doesn't work, then report them.*

In contrast, some respondents emphasized the importance of **protecting flaggers** (N = 60) to ensure fairness. They stressed the need for secure flagging processes that shield the identity of flaggers and prevent potential repercussions by authors of flagged posts. Some participants expressed that this confidentiality is also crucial in minimizing biases during content review, i.e., moderators should review flagged posts without knowing who flagged them. Alongside user protection, considerations for safeguarding **free speech** (N = 23) were also prominent. While acknowledging the role of flagging in community moderation, users cautioned against overly restrictive flagging practices that could stifle online expression. P555 wrote:

> *People need to understand that everyone has a right to their own opinions.*

6 Discussion

We began this study with the goal of examining how flagging mechanisms should be designed to enhance fairness perceptions among flaggers. Our statistical analysis of responses from a large-scale survey experiment shows that including posting guidelines and a text box for feedback within flag implementations helps enhance users' fairness perceptions, whereas offering classification schemes or providing information about whether the flag reviewer is a human or a bot does not significantly influence users' attitudes. Our qualitative analysis of open-ended responses shows that users feel concerned about reviewers' biases, desire flagging systems to support greater expressivity, demand timely notifications and explanations of flag outcomes, wish to see information about flag processing at the system-level, and expect platforms to prevent flag abuse.

These results contribute empirically informed guidance on how social media platforms should design different components of their flagging interfaces and how these design choices could impact users' attitudes toward flagging. We document the key information and security needs of flaggers and offer insights for how platforms could address them. We show that users' engagement with flags triggers a range of sociopolitical concerns regarding platforms' responsibilities, freedom of speech, algorithmic evaluation, safety against online harms, and privacy. Our study design also provides a methodological framework that others can adopt to evaluate the effectiveness of new components and affordances that seek to address users' reporting needs.

Prior empirical research on enacting fairness in content moderation [40, 42, 61, 66, 86] has largely focused on how moderation decisions are administered and communicated to moderated users. In contrast, our study foregrounds the experiences of flaggers and highlights their perceptions of procedural fairness at the *during reporting* stage of the flagging pipeline. Because flaggers provide indispensable labor for reporting systems [12, 25], ensuring their trust and ongoing participation is essential for sustaining a successful platform governance. Prior work also shows that enhancing procedural fairness in moderation can increase users' trust [70], even when moderation outcomes are unfavorable [86], and can encourage continued community participation [40]. By examining flaggers' perceptions of procedural justice in flag implementations, our findings identify

design considerations that can strengthen their trust and support their sustained engagement. Notably, we show that from the flaggers' perspective, each aspect of flagging mechanisms—the procedural elements available during flagging, flag review criteria, flagging outcomes, and how they are communicated—require greater consistency, transparency, and support for user expression. Attending to these user needs may lead to clearer, better-justified, more frequent, and higher-quality flags, thereby improving the efficiency of the moderation pipeline as a whole.

In designing this study, we did not ground our questions in a specific social media site to increase the generalizability of our results. Thus, our findings represent a baseline understanding of how users perceive flagging mechanisms regardless of their prior experiences with moderation or the platform's reputation. Our platform-agnostic design represents one end of a deliberate trade-off between achieving experimental control versus ecological validity that researchers must make in studying widely available moderation affordances like flagging. However, real-world flagging unfolds in environments where trust in platform varies widely, and users hold distinct expectations for different platforms. Moreover, in real settings, users' contextual ties with the flagged content may heighten their emotional involvement and render procedural details of the flag mechanism more consequential. Thus, platform-specific research that builds upon our work, especially the inquiries that longitudinally examine how users naturally interact with flags in their daily social media use, would offer further valuable insights.

Below, we discuss our design insights, elaborate upon our theoretical contributions regarding flagging as a content moderation mechanism, and suggest promising avenues for further research. We also present Figure 5, which offers the conceptual framework that guided our study design, maps how specific implementations of flag components shape users' fairness perceptions based on our empirical analysis, and summarizes our key findings. This framework serves as an overview that guides the following discussion.

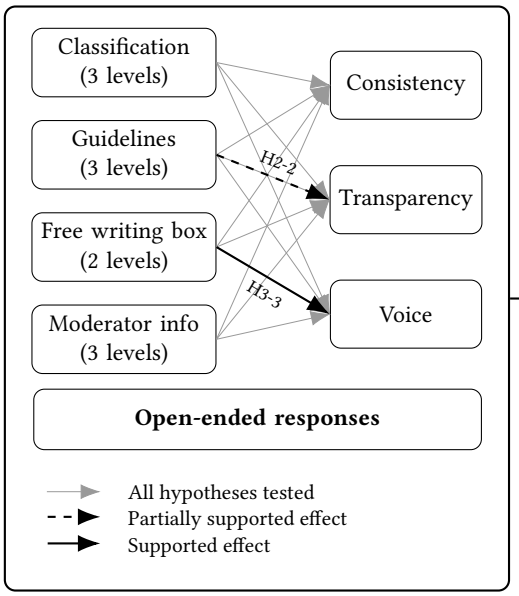
6.1 Enhancing Users' Perceptions of Being Heard and Allowing Detailed Expression

6.1.1. Incorporating a Text Box in Flagging Mechanisms

Our analysis reveals that flaggers' fairness perceptions, especially regarding having a voice in the content moderation process, improve when they have an opportunity to express in detail their objections to the post during the flagging process. Specifically, our quantitative analysis demonstrates that the availability of a text box, which lets users articulate their thoughts in their own words within the flagging mechanism, significantly enhances their sense of being heard (Sec. 5.1.3). Contrary to our hypothesis *H3-1* (Sec. 3.3), adding this text box to flag designs does not come at the cost of reducing users' perceived consistency (Sec. 5.1.3). This finding differs from prior work on designing appeal mechanisms to reverse moderation decisions, which showed that giving users an opportunity to submit text-based appeals does not significantly improve perceived voice or related fairness perceptions of moderation outcomes [86].

This suggests a crucial design implication: the timing or the context in which the user voice is solicited matters. In post-hoc moderation appeal settings, expression (by an already moderated user) occurs after a punitive decision has already been made, within a highly constrained jurisdictional frame that limits perceived voice. By contrast, flagging enables users to define the problem and articulate their own norms to inform the yet-to-begin decision-making process, and this context affords greater expressive freedom. Indeed, our findings indicate that treating users as early contributors to the decision-making process substantially strengthens their perceived voice.

Moreover, by positioning users as active agents within the moderation process, a text box could also enhance *outcome* fairness by providing flag reviewers with additional context to fairly evaluate flagged posts. Flag outcomes that are accompanied by explanations regarding decision-making [40,



Key quantitative findings

- Users shown (simple or detailed) guidelines perceive higher **transparency** than those without guidelines.
- Users shown a text box perceive higher **voice** than those without a box.

Key qualitative findings

- Users want more information about how flag processing occurs at the system level.
- Users desire greater expressivity, e.g., selecting multiple flagging categories and labeling norm-violating portions of flagged posts.
- Users feel concerned about potential biases and domain expertise of flag reviewers.
- Users desire educational initiatives on effective flagging and diverse channels (e.g., chat, phone calls) for reporting.
- Users would appreciate platform-hosted spaces that facilitate norm negotiation and sensemaking around flagging experiences.
- Users expect timely notifications, review status tracking, and explanations of flag outcomes.
- Users value both ex-ante harm prevention measures and protections for free speech when punitive tools are used.
- Users expect platforms to implement safeguards against flag abuse.

Fig. 5. On the left, this study's conceptual framework is shown, mapping how we hypothesized the relations between flagging design components and fairness perceptions, and highlighting instances when we rejected the null hypotheses. On the right, the key analysis results are summarized.

42] and how flaggers' voice was considered could improve flaggers' and flagged users' perceptions of both procedural and outcome fairness.

We found that a text box increases the cognitive load on users when flagging a post (Sec. 5.1.5), which may reduce the mechanism's overall usability. Although we did not find a direct association between this feature and users' intention to flag again, prior research indicates that mental fatigue during the flagging process can lead to negative user experiences [26]. Therefore, platforms should carefully consider adopting this feature since it may limit user participation. One way for flag mechanisms to reduce unwarranted cognitive load could be to make user input in this text box optional and clarify that it is to be used only if flaggers want to provide additional necessary context.

Further, we recognize that a rigorous implementation of this feature requires platform moderators to carefully consider user-submitted text when making their moderation decisions, which presents challenges of scale [27, 30]. The results of our interaction effects suggest a likely compromise: we found that when flagging mechanisms do not offer a text box, the ability to specify rule violation through a classification scheme becomes significantly more crucial to satisfying users' voice needs (Sec. 5.1.5). Therefore, we suggest that if platforms are concerned about the negative

impact of a text box on usability or cannot afford to implement one due to labor constraints, they should at least implement a robust classification scheme for rule violations (or another analogous feature that lets users adequately specify their post objections) in their flagging mechanisms.

6.1.2. Establishing Mechanisms to Track Flag Status

Our qualitative analysis underscores users' desire not only to voice their objections to the flagged post but also to ensure that their flag request has been successfully submitted, that they can monitor the flag processing status (Sec. 5.2.4), and that their flags receive timely decision-making by flag reviewers (Sec. 5.2.3). This emphasizes users' need for platforms to take their input in the moderation process seriously and to address their concerns without undue delays.

We echo participants' suggestion that platforms implement a streamlined tracking system that allows users to monitor the review progress of submitted flags. Such post-flagging interfaces could connect their design goals to values important to end-users, such as reliability, compassion, and social responsibility [87]. They could also offer affordances such as the ability to "unflag" a post or submit additional evidence after flagging. Previous HCI research efforts to design for contestability in moderation procedures [86, 87] could serve as a blueprint for taking a user-centered approach to building such flag-tracking systems. Some participants expressed a desire to provide feedback on their flagging experiences (Sec. 5.2.2). This suggests that offering a chance to be heard on the entire flagging process (i.e., beyond their objections to individual flagged posts) may influence users' fairness perceptions. Therefore, platforms could deploy a feedback system to collect recommendations for improvements and regularly evaluate the feedback received to update their current practices.

6.1.3. Widening Flags' Vocabulary of Complaint

Prior theoretical analysis of flagging mechanisms emphasized that flags offer "a narrow vocabulary of complaint" and do not account for the many complex reasons that people might choose to flag or let users specify their degree of concern with the flagged posts [12]. Our quantitative analysis did not find a relationship between the granularity of classification schemes and fairness perceptions (Sec. 5.1.1). However, we found that offering a classification scheme (either simple or detailed) enhances voice perceptions when a text box is not provided (Table 11) and providing this scheme does not impact usability (Sec. 5.1.5).

These partial findings suggest that designing classification schemes warrants further investigation, particularly in light of the sensitivity of our manipulation. Our operationalization distinguished simple from detailed schemes by adding a corresponding subcategory for each primary flag category. Yet, it is possible that the presence (or absence) of such subcategories does not matter much to end-users. If users associate granularity with richer design features—for example, selecting from a wide range of custom tags—our subcategory-based manipulation, although modeled on real-world implementations, may not have been psychologically salient enough to shift fairness perceptions. Moreover, the experimental task required participants to select one post from a set of seven that they already wished to report. Because the choice itself focuses participants on content they found most objectionable, classification granularity may have had limited influence on fairness perceptions when the violation category was already evident. Overall, these results indicate that asking users to categorize their objections using a classification scheme remains fundamentally a constricted way to submit reports from users' perspectives—and adding submenus to these schemes does not aid in addressing users' expression and fairness needs (Sec. 5.1.1).

Our qualitative results offer some clues on how platforms could expand flags' narrow vocabulary of complaint. Our participants suggested that flagging mechanisms deploy a classification scheme that allows the selection of multiple rule violations, an ability to highlight portions of the flagged post that violate platform rules, and a way to rate how severely inappropriate the post is (Sec. 5.2.2).

Studies that experiment with such additional innovations in flag designs (e.g., finding optimal taxonomy on which flaggers rate a violation's severity) would be valuable. Participants' demand for such features indicates that users feel a need to exert a greater level of voice and control over the moderation procedures, echoing previous research findings on users interacting with personal moderation [44] and community management tools [41].

While users desire expressive flagging options, it is crucial that platforms provide such options only when they can effectively incorporate that expression. For instance, allowing flaggers to provide free-text responses may enhance their sense of having a voice (perceived fairness) during flag submission, but it does not improve *actual* procedural fairness if the flag reviewers lack the ability to process that input when making moderation decisions. Such ineffectiveness could, in the long run, backfire and further reduce users' trust in the moderation process.

In sum, enacting flagging solutions that let users clarify the post's context and offer greater flexibility than having to shoehorn complex feelings into a single category selection [12] would enhance flaggers' fairness perceptions.

6.2 Incorporating Transparency in Review Criteria, Reviewers, and Review Outcomes

6.2.1. Integrating Posting Guidelines in Flagging Mechanisms

Our statistical analyses show that incorporating posting guidelines into the flagging process enhances participants' perceptions of fairness (Sec. 5.1.2). Specifically, providing these guidelines improves users' perceptions of moderation transparency, and providing additional information—such as examples of rule-violating posts in these guidelines—further contributes to users recognizing the flagging process as more transparent. This aligns with prior research on the design of personal moderation interfaces [44], where including examples of rule-violating posts enhanced users' perceptions of control over the moderation process. We also found that contrary to our hypothesis *H2-3* (Sec. 3.2), integrating posting guidelines does not compromise users' perceptions of voice (Sec. 5.1.2), and neither does it raise users' cognitive burdens (Sec. 5.1.5)—these results further incentivize showing posting guidelines to flaggers.

However, most mainstream platforms currently *do not* include posting guidelines in their flagging designs; we found that, currently, only Facebook and Instagram link users to these guidelines. We suggest that platforms list or link posting guidelines in the flagging mechanism to enhance their users' understanding of the flag review process and improve their transparency perceptions.

6.2.2. Offering Information About Flag Reviewers

Our quantitative analysis shows that fairness perceptions of flagging procedures are not influenced by information about the moderator type (a human, a bot, or no information) (Sec. 5.1.4), which aligns with prior research findings [29, 65, 69]. However, our qualitative findings suggest that participants were not simply reacting to whether a reviewer was human or AI; instead, they expressed interest in more granular information about reviewers' qualifications. We found that disclosing moderators' professional characteristics, e.g., information about their experiences and skills, such as the training they receive or their expertise in specific subjects, can foster greater trust in the flag review procedures (Sec. 5.2.1). Because these findings are exploratory, future work is needed to evaluate whether such information meaningfully affects fairness perceptions in practice and how it can be presented while still respecting moderators' privacy.

Our qualitative findings also highlight users' preference for involving multiple reviewers in content moderation (Sec. 5.2.1). Although our quantitative analyses did not test for multiple reviewers, participants suggested that using more than one moderator or combining human and bot reviewers could enhance fairness in the flag review process. This aligns with Fan and Zhang's finding

that group-based moderation improves perceptions of fairness [21] and Katsaros et al.'s observation that users prefer systems that combine humans with algorithmic decisions [47], supporting the need for hybrid review models. Consistent with prior work [20, 70], users' preference for multiple moderators appears to stem from concerns about potential biases when a single moderator makes decisions. Thus, even though deploying multiple moderators could be resource intensive, our participants' responses underscore the importance of exploring mechanisms that incorporate diverse perspectives and substantively safeguard against reviewer biases.

6.2.3. Reforming Post-flag Submission Steps

Participants emphasized the need for greater transparency in the flag review process and the disclosure of information about post-flag submission steps. This involves revealing the review criteria, the specifics of each step in the flag review, the expected timeline, and the information visible to reviewers about flaggers (Sec. 5.2.4).

Additionally, our qualitative insights show users' concerns that the flag review process may not adequately address the problem of malicious [31] or organized [12] flagging, i.e., flagging of content that does not violate platform guidelines (Sec. 5.2.5). Since a flag may not accurately indicate the post's actual inappropriateness [52], e.g., some people may exploit flags as a form of 'digilantism' or politically motivated extrajudicial practice [38], users feel concerned about unjust sanctions against norm-complying content. We suggest that platforms assuage such concerns by informing users about the measures they take to prevent malicious users from abusing flagging. For example, they could specify their procedures for sanctioning users who repeatedly engage in false flagging. Further, in cases where flagged posts fall into borderline or ambiguous categories, platforms could incorporate additional verification steps—such as secondary review or escalation to a specialized moderation tier—to increase confidence in the legitimacy of the final decisions.

6.3 Supporting Flaggers with Different Technological Competencies and Diverse Perspectives

6.3.1. Improving Flagging Visibility and Accessibility

Participants' open-ended responses indicate a desire for flagging systems to have greater accessibility and improved usability to empower more users to report inappropriate posts (Sec. 5.2.2). Our analysis also surfaced a need to extend awareness about flags, e.g., by educating users on how to flag objectionable content effectively, why such flagging is important, and how their feedback is processed. Platforms themselves can play an important role in such educational initiatives. As noted by Naab et al. [67], platforms often fail to encourage user engagement by not providing clear, accessible information on flagging uncivil posts, highlighting the need for more visible guidance on flagging. Further, it is vital to promote user participation through clear descriptions of flagging steps, offering incentives, and ensuring the simplicity and convenience of the flagging process. Additionally, establishing diverse channels for reporting inappropriate content, such as phone calls, emails, and chats, could help users with different technological competencies and preferences to flag in ways they find intuitive and accessible.

6.3.2. Hosting Discussion Forums for Flagging

As participants suggested, platforms could also create special forums for discussions centered around flagging (Sec. 5.2.2). For example, such forums could host conversations about whether certain controversial posts should be flagged or how platforms' moderation policies and flagging classification schemes do not accommodate certain norm violations. They could also offer users a converging space to discuss their flagging experiences, e.g., forum members could share their

flagging history regarding the outcomes of posts they previously flagged. Such spaces could allow users to appreciate diverse perspectives regarding content moderation and develop a shared understanding of how platforms respond to flagging efforts.

One challenge with hosting such forums is that a narrow set of influential individuals with strong viewpoints or even bad actors may unduly influence discussions about flagging norms, e.g., conversations about whether certain posts should be flagged [23]. Therefore, platforms should carefully design such gatherings in collaboration with a diverse set of stakeholders, state their purpose clearly, and keep them well moderated.

6.4 Addressing Online Harms Holistically

6.4.1. Enacting Outcome Fairness in Flagging Mechanisms

Our study investigated how different components of the flagging system influence users' perceptions of procedural fairness. Procedural fairness, as studied by several researchers [79, 85, 86], is characterized by its non-consequential nature and concerns users' experience of the procedural steps [59]. However, our qualitative analysis revealed that many participants link fairness in flagging procedures to the future decision outcomes (Sec. 5.2.3).

Specifically, participants observed that flagging mechanisms can enhance their fairness by providing clear information about outcomes accompanied by detailed reasoning for outcome decisions. Some participants also indicated a need for a rebuttal system that lets flaggers challenge flagging outcomes (Sec. 5.2.2). Others were curious to see comprehensive statistics about the regulation of flagged content, including the proportion of submitted posts that are flagged and the ratio of flagged posts sanctioned within specific timeframes (Sec. 5.2.4). These suggestions indicate that users conceptualize fairness of flagging mechanisms in a holistic manner, especially attending to flagging outcomes and how they are communicated as well as platform-wide measures associated with flagging.

We recommend that platforms include relevant system-wide descriptive information about the use of flags in their transparency reports. This could include the number of flags submitted under each flagging category, alongside a summary of their review outcomes. It would be valuable to learn the extent of false (or malicious) flags that platforms detect on their site, and the origins (e.g., flaggers' countries) and targets (e.g., how many of the accounts falsely flagged belong to politicians, public figures, etc.) of such actions. Platforms could also reveal how flag reviews continue to shape other elements of the moderation ecosystem, e.g., how they are used to enhance automated decision-making.

6.4.2. Encouraging Norm Compliance Among Flagged Users

Beyond concerns about whether the flagged post was sanctioned [12], flaggers may be invested in how effective their flagging efforts are at preventing further norm violations by the flagged users. Indeed, our findings show that some users prefer educating rule-violating users rather than merely taking punitive measures against them, such as removing their content after flag review (Sec. 5.2.5).

Therefore, platforms should consider investing in educational measures, such as helping the authors of flagged posts better understand the community rules and how to adhere to them. As prior research shows, moderator messages that explain to the sanctioned users *why* their posts are removed help improve their attitudes and future behaviors [40, 42]. Highlighting, rewarding, and incentivizing desirable behaviors also reinforce constructive contributions [10, 54]. Further, platforms may strengthen ex ante moderation measures [32], such as surfacing posting guidelines while a user is writing a post or using AI-based tools to warn users when their post draft is likely

to be sanctioned [39, 49]. In line with findings from Zhang et al. [93], some of our respondents preferred directly communicating with rule-violators over reporting them to persuade norm compliance. Such measures may reduce the burden of flagging on regular users and flag review on moderators.

6.4.3. Supporting Free Speech Values

Our qualitative analysis also highlighted that upholding free speech values in moderation mechanisms is critical for many end-users (Sec. 5.2.5). Some participants stressed that using flags as a tool to induce post removals may violate free speech principles. As prior research points out, free speech proponents support the use of personal moderation tools like muting and blocking, especially when compared to platform-enacted moderation, because these tools affect only the configuring user's newsfeed [44, 45]. Thus, platforms might consider informing users about these alternative options to address content-based harms [43] and clarifying their distinct affordances when they attempt to use the flagging tools.

6.5 Limitations and Future Work

We captured our participants' demographic details directly from Lucid's pre-collected data, which does not allow reporting of non-binary gender identities. Future studies using Lucid Theorem can capture this information using a separate question in the survey. Additionally, we measured each of the three procedural fairness aspects with a single-item indicator designed to efficiently capture participants' immediate reactions to our design manipulations. Future studies can further increase the reliability of results by employing validated multi-item scales to capture the multifaceted nature of each fairness dimension.

Our experiment evaluated participants' flagging of a single inappropriate post selected from a set of seven rule-violating examples. This small sample cannot represent the wide range of content-based harms that could trigger flagging. Additionally, different users might experience different levels of outrage in response to the same post, and this outrage level could be an important influence on their cognitive evaluations and fairness perceptions of flags—this could be a valuable metric to assess in future research. On the other hand, a preliminary analysis we conducted showed no significant differences in fairness perceptions based on the initially selected examples, supporting the robustness of our findings.

Our use of a between-subjects experimental design meant that each participant experienced only one flagging implementation. This design choice precluded collecting participant feedback on why many of the specific interface differences we tested did not yield significant shifts in fairness perceptions. Future studies that deploy within-subjects designs could offer valuable insights into how users perceive and compare particular differences in flag designs.

Moreover, users' behaviors may vary in their daily social media use as they encounter multiple instances of norm violations. It is possible that flagging multiple posts (instead of a single post, as tested in this study) could give users more experience and have them develop stronger stances on flags' procedural fairness. While the use of highly offensive stimuli in our survey creates a strong setting for evaluating flag mechanisms, less extreme content might shape flagging needs and actions differently. Therefore, longitudinal or in situ analyses that examine how users interact with flagging tools in their day-to-day settings would be valuable to further inform the tradeoffs between incorporating fairness and reducing cognitive burden.

7 Conclusion

This paper examines how flag components that provide different types of information about the flag review process shape users' attitudes toward flagging. Our analysis shows that including posting guidelines in flag designs enhances users' transparency perceptions and offering a text box improves their voice perceptions. We found that users desire flag mechanisms to support greater expressivity, timely notifications of flag review updates, increased visibility into flag review procedures and reviewers, and stronger protections against flag abuse. We discuss how innovations in flagging systems, such as establishing diverse channel for reporting and offering support for highlighting the severity of rule violations, could better support end-users' fairness demands. This investigation demonstrates how the design and policy choices made in the implementation of flagging infrastructures deeply shape users' daily experiences of social media use and address (or fail to address) their vital needs to combat content-based harms. We call for future studies to deploy similar user-centered approaches and social justice orientations to improve the current practices of platform governance.

Acknowledgments

We thank the anonymous reviewers for their generous feedback that helped strengthen this manuscript. This research was supported by the National Science Foundation award 2329394.

References

[1] Carolina Are. 2023. Flagging as a silencing tool: Exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok. *New Media & Society* (2023), 14614448241228544.

[2] Shubham Atreja, Libby Hemphill, and Paul Resnick. 2023. Remove, reduce, inform: what actions do people want Social Media platforms to take on potentially misleading content? *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.

[3] Ben Bradford, Florian Grisel, Tracey L Meares, Emily Owens, Baron L Pineda, Jacob Shapiro, Tom R Tyler, and Danieli Evans Peterman. 2019. Report of the Facebook data transparency advisory group. *Yale Justice Collaboratory* (2019).

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[5] Albert Breton, Gianluigi Galeotti, Pierre Salmon, Ronald Wintrobe, et al. 2007. *The economics of transparency in politics*. Ashgate Aldershot.

[6] W Chan Kim and Renée Mauborgne. 1998. Procedural justice, strategic decision making, and the knowledge economy. *Strategic management journal* 19, 4 (1998), 323–338.

[7] 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.

[8] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 32 (nov 2018), 25 pages. doi:10.1145/3274301

[9] Wallace Chipidza and Jie Yan. 2022. The effectiveness of flagging content belonging to prominent individuals: The case of Donald Trump on Twitter. *Journal of the Association for Information Science and Technology* 73, 11 (2022), 1641–1658.

[10] Frederick Choi, Charlotte Lambert, Vinay Koshy, Sowmya Pratipati, Tue Do, and Eshwar Chandrasekharan. 2024. Creator Hearts: Investigating the Impact Positive Signals from YouTube Creators in Shaping Comment Section Behavior. *arXiv preprint arXiv:2404.03612* (2024).

[11] Juliet Corbin and Anselm Strauss. 2015. *Basics of qualitative research*. Vol. 14. sage.

[12] 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.

[13] Elina Dale, Elizabeth F Peacocke, Espen Movik, Alex Voorhoeve, Trygve Ottersen, Christoph Kurowski, David B Evans, Ole Frithjof Norheim, and Unni Gopinathan. 2023. Criteria for the procedural fairness of health financing decisions: a scoping review. *Health Policy and planning* 38, Suppl 1 (2023), i13.

[14] Dipto Das, Carsten Østerlund, and Bryan Semaan. 2021. " Jol" or" Pani"?: How Does Governance Shape a Platform's Identity? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[15] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[16] 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. https://www. brennancenter. org/our-work/research-reports/double-standards-socialmedia-content-moderation* (2021).

[17] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. 2018. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[18] Denitsa Dineva and Jan Breitsohl. 2022. Managing trolling in online communities: an organizational perspective. *Internet Research* 32, 1 (2022), 292–311.

[19] Peter Drahos. 2017. *Regulatory theory: Foundations and applications*. ANU Press.

[20] Stefanie Duguay, Jean Burgess, and Nicolas Suzor. 2020. Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence* 26, 2 (2020), 237–252.

[21] Jenny Fan and Amy X Zhang. 2020. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[22] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. 2020. Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.

[23] Dongfang Gaozhao. 2021. Flagging fake news on social media: An experimental study of media consumers' identification of fake news. *Government Information Quarterly* 38, 3 (2021), 101591.

[24] Sarah A Gilbert. 2020. " I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

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

[26] Tarleton Gillespie. 2018. Regulation of and by platforms. *The SAGE handbook of social media* (2018), 254–278.

[27] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234.

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

[29] João Gonçalves, Ina Weber, Gina M Masullo, Marisa Torres da Silva, and Joep Hofhuis. 2021. Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. *new media & society* (2021), 14614448211032310.

[30] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.

[31] Rachel Griffin. 2022. The sanitised platform. *J. Intell. Prop. Info. Tech. & Elec. Com. L.* 13 (2022), 36.

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

[33] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[34] Ivar A. Hartmann. 2020. A new framework for online content moderation. *Computer Law & Security Review* 36 (2020), 105376. doi:10.1016/j.clsr.2019.105376

[35] Andrew F Hayes. 2020. *Statistical methods for communication science*. Routledge.

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

[37] Nataliya V Ivankova and John W Creswell. 2009. Mixed methods. *Qualitative research in applied linguistics: A practical introduction* 23 (2009), 135–161.

[38] Emma A Jane. 2017. 'Dude… stop the spread': antagonism, agonism, and# manspreading on social media. *International Journal of Cultural Studies* 20, 5 (2017), 459–475.

[39] Sarah Jerasa and Sarah K Burriss. 2024. Writing with, for, and against the algorithm: TikTokers' relationships with AI as audience, co-author, and censor. *English Teaching: Practice & Critique* 23, 1 (2024), 118–134.

[40] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.

[41] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[42] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[43] 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

[44] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.

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

[46] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2023. A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–34.

[47] Matthew Katsaros, Jisu Kim, and Tom Tyler. 2024. Online content moderation: does justice need a human face? *International Journal of Human–Computer Interaction* 40, 1 (2024), 66–77.

[48] Matthew Katsaros, Tom Tyler, Jisu Kim, and Tracey Meares. 2022. Procedural justice and self governance on Twitter: Unpacking the experience of rule breaking on Twitter. *Journal of Online Trust and Safety* 1, 3 (2022).

[49] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 2022. Reconsidering tweets: Intervening during tweet creation decreases offensive content. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 477–487.

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

[51] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–21.

[52] 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*. 1–12.

[53] Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

[54] Charlotte Lambert, Frederick Choi, and Eshwar Chandrasekharan. 2024. "Positive reinforcement helps breed positive behavior": Moderator Perspectives on Encouraging Desirable Behavior. (2024).

[55] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[56] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the Monetary Value of Online Volunteer Work. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 596–606. doi:10.1609/icwsm.v16i1.19318

[57] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the appeal? Perceptions of review processes for algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

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

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

[60] Renkai Ma, Yao Li, and Yubo Kou. 2023. Transparency, Fairness, and Coping: How Players Experience Moderation in Multiplayer Online Games. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[61] Renkai Ma, Yue You, Xinning Gui, and Yubo Kou. 2023. How Do Users Experience Moderation?: A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 278 (oct 2023), 30 pages. doi:10.1145/3610069

[62] Jessica Maddox and Jennifer Malson. 2020. Guidelines without lines, communities without borders: The marketplace of ideas and digital manifest destiny in social media platform policies. *Social Media+ Society* 6, 2 (2020), 2056305120926622.

[63] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.
[64] Jisu Kim Matthew Katsaros and Tom Tyler. 2024. Online Content Moderation: Does Justice Need a Human Face? *International Journal of Human–Computer Interaction* 40, 1 (2024), 66–77. doi:10.1080/10447318.2023.2210879 arXiv:https://doi.org/10.1080/10447318.2023.2210879
[65] Maria D Molina and S Shyam Sundar. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication* 27, 4 (2022), zmac010.
[66] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
[67] Teresa K Naab, Anja Kalch, and Tino GK Meitz. 2018. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20, 2 (2018), 777–795.
[68] Michael Nycyk. 2016. Enforcing community guidelines in web-based communities: the case of flame comments on YouTube. *International Journal of Web Based Communities* 12, 2 (2016), 131–146.
[69] 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.
[70] Christina A Pan, Sahil Yakhmi, Tara P Iyer, Evan Strasnick, Amy X Zhang, and Michael S Bernstein. 2022. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–31.
[71] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
[72] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.
[73] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 368 (oct 2021), 33 pages. doi:10.1145/3479512
[74] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from justice theories to support targets of online harassment. *new media & society* 23, 5 (2021), 1278–1300.
[75] Joseph Seering. 2020. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–28.
[76] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 111–125.
[77] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
[78] Donghee Shin and Yong Jin Park. 2019. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98 (2019), 277–284.
[79] Jason Sunshine and Tom R Tyler. 2003. The role of procedural justice and legitimacy in shaping public support for policing. *Law & society review* 37, 3 (2003), 513–548.
[80] Nicolas P Suzor. 2019. *Lawless: The secret rules that govern our digital lives*. Cambridge University Press.
[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] Sue Tait. 2008. Pornographies of Violence? Internet Spectatorship on Body Horror. *Critical Studies in Media Communication* 25, 1 (March 2008), 91–111. doi:10.1080/15295030701851148
[83] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2022. (In) visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *new media & society* (2022), 14614448221109804.
[84] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. 2022. "It's Common and a Part of Being a Content Creator": Understanding How Creators Experience and Cope with Hate and Harassment Online. 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 121, 15 pages. doi:10.1145/3491102.3501879
[85] Tom R Tyler. 2006. Psychological perspectives on legitimacy and legitimation. *Annu. Rev. Psychol.* 57 (2006), 375–400.
[86] 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. *Proceedings of the ACM on human-computer interaction* 4, CSCW2 (2020), 1–22.

[87] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability for content moderation. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (2021), 1–28.
[88] Camilla Jung Westermann and Michele Coscia. 2022. A potential mechanism for low tolerance feedback loops in social media flagging systems. *Plos one* 17, 5 (2022), e0268270.
[89] Richard Ashby Wilson and Molly K Land. 2020. Hate speech on social media: Content moderation in context. *Conn. L. Rev.* 52 (2020), 1029.
[90] Magdalena Wojcieszak, Arti Thakur, João Fernando Ferreira Gonçalves, Andreu Casas, Ericka Menchen-Trevino, and & Miriam Boon. 2021. Can AI enhance people's support for online moderation and their openness to dissimilar political views? *Journal of Computer-Mediated Communication* 26, 4 (2021), 223–243.
[91] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
[92] Jing Zeng and D Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet* 14, 1 (2022), 79–95.
[93] Alice Qian Zhang, Kaitlin Montague, and Shagun Jhaver. 2023. Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Posts. *arXiv preprint arXiv:2309.06688* (2023).
[94] Andrew Zolides. 2021. Gender moderation and moderating gender: Sexual content policies in Twitch's community guidelines. *New Media & Society* 23, 10 (2021), 2999–3015.

A Survey Sample Description

Table 8 shows the demographic characteristics of the survey respondents. Our participants included 1,441 males and 1,495 females, and they had a mean age of 45. The majority of income brackets of our sample were between $25,000 and $49,999 (25.0%) and less than $25,000 (24.9%). Study participants were predominantly White (72.3%), followed by Black or African American (12.2%), and other ethnic groups (15.5%). Geographic distribution showed that respondents were predominantly from the South (37.6%), followed by the West (23.8%), the Northeast (20%), and the Midwest (18.6%). Daily social media usage was the most prevalent (67.3%), with only 12.3% of participants reporting less frequent than weekly use. Educational attainment varied widely: 33.9% attended some high school or less; 53.3% attended some college, including AD and BA; and 12.8% had graduate degrees.

B Fairness Perceptions for Different Flagged Posts

Table 9 shows variations in fairness perceptions for different selections of inappropriate posts to flag by survey respondents.

C Additional Analyses

Table 10 below presents the GLM analysis results, which indicate how different combinations of flagging components interact, affecting perceived consistency, transparency, and voice. Next, we present Table 11, which details our results on interaction effects of the classification scheme and a text box on perceived voice. Tables 12 and 13 summarize our results on how different flagging components influence usability. We comment on these results in sec. 5.1.5.

Table 8. Demographics of Survey Respondents.

Demographic Factor	Category	Number (%)
Gender	Male	1441 (48.6%)
	Female	1495 (50.4%)
Age	Range: 18-89 (Mean = 45)	-
Ethnicity	White	2123 (72.3%)
	Black or African American	359 (12.2%)
	Asian	152 (5.2%)
	Pacific Islander	254 (8.7%)
	American Indian or Alaska Native	48 (1.6%)
Hispanic, Latino, or Spanish Origin	Yes	377 (12.8%)
	No	2559 (87.2%)
Income	Less than $25,000	736 (25%)
	$25,000 to $49,999	738 (25.1%)
	$50,000 to $74,999	543 (18.5%)
	$75,000 to $124,999	555 (19%)
	$125,000 and above	336 (11.4%)
	Prefer not to answer	28 (1%)
Political affiliation	Democrat	1217 (41.5%)
	Republican	1090 (37.1%)
	Neutral	659 (21.4%)
Geographic region	South	1104 (37.6%)
	West	699 (23.8%)
	Northeast	587 (20%)
	Midwest	546 (18.6%)
Social media use frequency	Never	168 (5.7%)
	Once a week	193 (6.6%)
	2-3 times a week	291 (9.9%)
	4-6 times a week	308 (10.5%)
	Daily	1976 (67.3%)
Educational attainment	Some high school or less	996 (33.9%)
	Some college including AD, BA	1565 (53.3%)
	Master's degree or equivalent	281 (9.6%)
	Doctorate degree	72 (2.5%)

Table 9. The mean values of perceived consistency, transparency, and voice for each of the rule violation examples that participants choose to flag in the survey. Standard deviation (SD) values are shown in brackets.

Category	N	Mean (SD)		
		Consistency	Transparency	Voice
Mexicans come from an uncivilized ...	621	5.57 (1.45)	5.33 (1.31)	5.31 (1.49)
@Sean11 I hate all you F*** ...	272	5.50 (1.52)	5.39 (1.29)	5.33 (1.43)
Drinking bleach has been scientifically ...	694	5.68 (1.37)	5.42 (1.36)	5.35 (1.56)
Today is a great day! I ate 723 calories ...	165	5.09 (1.62)	4.98 (1.64)	5.01 (1.74)
Download the software for credit card ...	178	5.34 (1.57)	5.01 (1.49)	5.00 (1.60)
Omg just got tons of Bucks from here! ...	134	5.24 (1.56)	5.00 (1.41)	4.86 (1.68)
I have a masturbation video of @Janny12 ...	872	5.60 (1.39)	5.34 (1.37)	5.43 (1.51)
Total	**2,936**	**5.54 (1.45)**	**5.30 (1.38)**	**5.30 (1.55)**

Table 10. GLM Results, Indicating the Interaction Effects on Perceived Consistency, Transparency, and Voice.

Fairness aspect	Interaction between variables	SS	df	MS	F
Consistency	Classification	8.35	2	4.18	2.0
	Guidelines	3.26	2	1.63	.80
	Text box	-	1	-	-
	Moderator	1.00	2	.50	.24
	Classification × Guidelines	10.68	4	2.67	1.27
	Classification× Text box	8.53	2	4.27	2.02
	Classification × Moderator	4.72	4	1.18	.56
	Guidelines × Text box	4.89	2	2.44	1.16
	Guidelines × Moderator	3.95	4	.99	.47
	Text box × Moderator	6.71	2	3.35	1.59
	Classification× Guidelines × Text box	5.17	4	1.29	.61
	Classification× Guidelines × Moderator	27.54	8	3.44	1.63
	Classification × Text box × Moderator	1.63	4	.41	.19
	Guideline × Text box × Moderator	11.79	4	2.95	1.40
	Classification × Guidelines × Text box × Moderator	17.37	8	2.17	1.03
Transparency	Classification	3.53	2	1.77	.94
	Guidelines	**41.55**	**2**	**20.78**	**11.05*****
	Text box	2.15	1	2.15	2.14
	Moderator	1.71	2	.86	.46
	Classification × Guidelines	1.78	4	.44	.24
	Classification × Text box	4.17	2	2.09	1.11
	Classification × Moderator	8.27	4	2.07	1.10
	Guidelines × Text box	**15.89**	**2**	**7.95**	**4.22***
	Guidelines × Moderator	6.88	4	1.72	.92
	Text box × Moderator	9.13	2	4.57	2.43
	Classification× Guidelines × Text box	9.28	4	2.32	1.23
	Classification × Guidelines × Moderator	13.11	8	1.64	.87
	Classification × Text box × Moderator	13.71	4	3.43	1.82
	Guidelines × Text box × Moderator	7.91	4	1.98	1.05
	Classification × Guidelines × Text box × Moderator	12.69	8	1.59	.84
Voice	Classification	10.24	2	5.12	2.27
	Guidelines	6.42	2	3.21	1.43
	Text box	**404.06**	**1**	**404.06**	**179.48*****
	Moderator	1.89	2	.95	.42
	Classification × Guidelines	10.11	4	2.53	1.12
	Classification × Text box	**37.33**	**2**	**18.67**	**8.29*****
	Classification × Moderator	14.30	4	3.58	1.59
	Guidelines × Text box	4.66	2	2.33	1.04
	Guidelines × Moderator	1.41	4	.35	.16
	Text box × Moderator	1.63	2	.82	.36
	Classification × Guidelines × Text box	2.65	4	.66	.29
	Classification × Guidelines × Moderator	12.85	8	1.61	.71
	Classification × Text box × Moderator	3.41	4	.85	.38
	Guidelines × Text box × Moderator	5.58	4	1.39	.62
	Classification × Guidelines × Text box × Moderator	9.55	8	1.19	.53

* indicates $p < 0.05$ and *** indicates $p < 0.001$.

Table 11. Interaction Effects of Classification Scheme and Text Box on Perceived Voice.

Fairness aspect	Interaction between the two variables		MD	SE
Voice	No text box	**Simple - No classification**	**.34*****	**.10**
		Detailed - No classification	**.35*****	**.10**
		Detailed - Simple classification	.01	.10
	Text box is provided	Simple - No classification	-.05	.10
		Detailed - No classification	-.20	.10
		Detailed - Simple classification	-.15	.10

*** indicates $p < 0.001$.

Table 12. Results Summarizing Whether Different Choices of Flagging Components Impact Participants' Cognitive Burden.

Components	SS	df	MS	F	*t*
Classification	8.14	2	4.07	1.31	-
Guidelines	17.63	2	8.81	2.83	-
Text box	**-**	**2933.87**	**-**	**-**	**-3.03****
Moderator	4.01	2	2.00	0.64	-

Table 13. Results Summarizing Whether Different Choices of Flagging Components Impact Participants' Future Use.

Components	SS	df	MS	F	*t*
Classification	.91	2	.46	.19	-
Guidelines	.12	2	.06	.03	-
Text box	-	2930.17	-	-	-1.07
Moderator	8.40	2	4.20	1.77	-

** indicates $p < 0.01$.

In the above tables, ANOVA results are presented for 'Classification,' 'Guidelines,' and 'Moderator' components and t-test results are presented for the 'Text Box' component.