Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor

SHAGUN JHAVER, Rutgers University, New Brunswick, NJ, USA
ALICE QIAN ZHANG, University of Minnesota, Minneapolis, MN, United States
QUANZE CHEN, University of Washington, Seattle, WA, USA
NIKHILA NATARAJAN, Rutgers University, New Brunswick, NJ, USA
RUOTONG WANG, University of Washington, Seattle, WA, USA
AMY ZHANG, University of Washington, Seattle, WA, USA

Social media platforms moderate content for each user by incorporating the outputs of both platform-wide content moderation systems and, in some cases, user-configured personal moderation preferences. However, it is unclear (1) how end users perceive the choices and affordances of different kinds of personal content moderation tools, and (2) how the introduction of personalization impacts user perceptions of platforms' content moderation responsibilities. This paper investigates end users' perspectives on personal content moderation tools by conducting an interview study with a diverse sample of 24 active social media users. We probe interviewees' preferences using simulated personal moderation interfaces, including word filters, sliders for toxicity levels, and boolean toxicity toggles. We also examine the labor involved for users in choosing moderation settings and present users' attitudes about the roles and responsibilities of social media platforms and other stakeholders towards moderation. We discuss how our findings can inform design solutions to improve transparency and controllability in personal content moderation tools.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Shagun Jhaver, Alice Qian Zhang, Quanze Chen, Nikhila Natarajan, Ruotong Wang, and Amy Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *J. ACM* xx, x, Article xx (x 2023), 33 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

At the Game Developers Conference in 2021, Intel unveiled Bleep, a new AI-powered tool for users to filter out online abuse in video game voice chat.[1] As shown in screenshots of the demo in Fig. 1, the tool listed commonly encountered categories of abuse, like 'Aggression' and 'Misogyny,'

[1]https://youtu.be/97Qhj299zRM?t=1781

Authors' addresses: Shagun Jhaver, Rutgers University, New Brunswick, NJ, USA, shagun.jhaver@rutgers.edu; Alice Qian Zhang, University of Minnesota, Minneapolis, MN, United States, zhan6698@umn.edu; Quanze Chen, University of Washington, Seattle, WA, USA, cqz@cs.washington.edu; Nikhila Natarajan, Rutgers University, New Brunswick, NJ, USA, nn352@rutgers.edu; Ruotong Wang, University of Washington, Seattle, WA, USA, ruotongw@cs.washington.edu; Amy Zhang, University of Washington, Seattle, WA, USA, axz@cs.uw.edu.

0004-5411/2023/x-ARTxx $15.00
https://doi.org/XXXXXXX.XXXXXXX

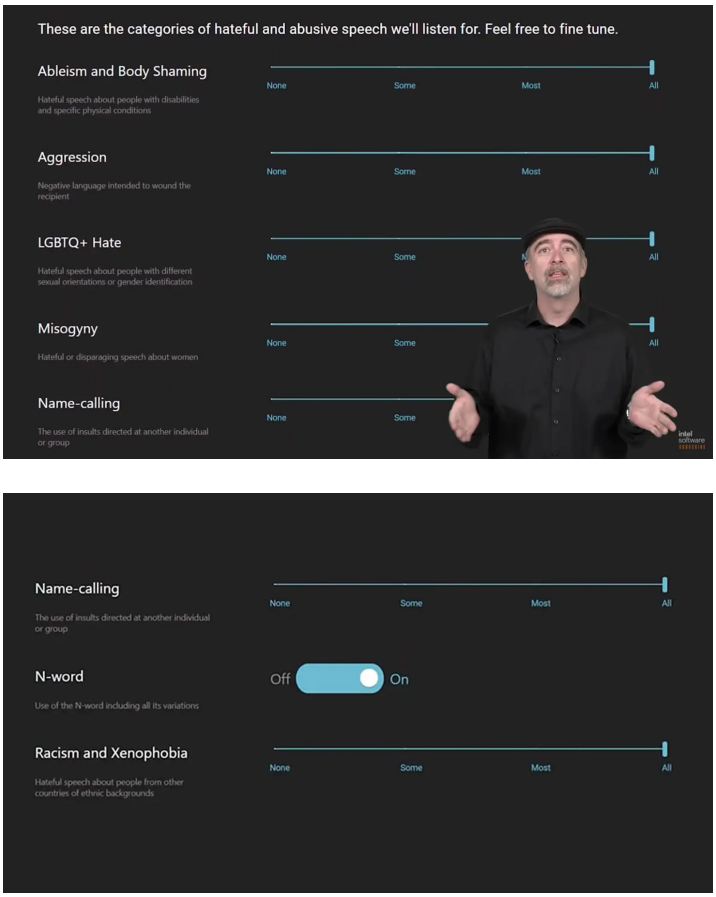


Fig. 1. Screenshots of Intel's implementation of a personal content moderation tool using sliders for different categories. During our interviews, we showed these images to our participants to elicit their suggestions.

each paired with a slider to control the quantity of each category that a user wants to hear [85]. Reactions were swift and heated across news and social media [21]. One article objected to how the tool "*pitch[es] racism, xenophobia, and general toxicity as settings that can be tuned up and down as though they were...control sliders on a video game*" [29]. Other people argued that such a tool would be beneficial; for example, one Twitter user said, paraphrasing: *"...as a trans woman gamer I pretty much always hear transphobia and misogyny online, but with this I can just bleep those words instead of muting people manually over and over..."* Still others questioned the appropriateness of asking users to configure moderation settings, with a Twitter user saying: *" ...WTF kind of dystopic insanity is this control panel? Why would the onus be on the user to do the filtering?"*

The mix of responses highlights some of the difficult questions that arise when tools for personal configuration of moderation settings are introduced on a social media platform. These tools have the potential to address online speech harms in a more user-initiated and personalized way. However, the rich debate around the Intel demo shows that users have different opinions on and preferences for such tools.

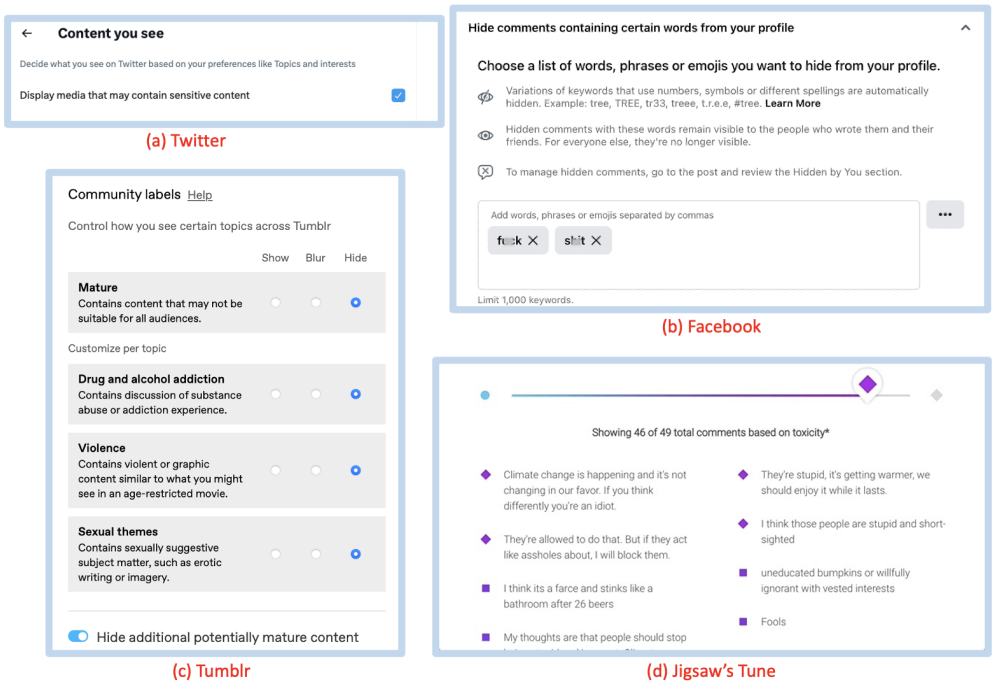


Fig. 2. Examples of Personal Content Moderation Tools on (a) Twitter, (b) Facebook, (c) Tumblr, and (d) Jigsaw Tune.

Since the mid-2000s, social media platforms' popularity has expanded dramatically worldwide [7, 101, 104]. User-generated content (UGC) can be empowering, especially for those belonging to vulnerable or marginalized groups, as it allows users to dictate what content is created [87]. However, the rules around what content is acceptable on a platform versus not continue to be narrowly shaped by the cultural norms of Silicon Valley, where most big platforms hosting UGC are located [33]. Given the normative differences across cultures [57] and communities [112], including even those that are geographically co-located [113], a one-size-fits-all solution to shaping online content would not be able to serve the disparate needs of millions of end users [56].

Recognizing the inevitable conflicts regarding platform-wide content moderation, some industry leaders, scholars, and activists have called for an approach that we refer to as **personal content moderation**. We define personal content moderation as *a form of moderation in which users can configure or customize some aspects of their moderation preferences on social media based on the content.* Recently, platforms have also begun to experiment with and offer such tools. They are 'personal' in that every user can configure them differently, and a user's configuration applies only to their own account. In addition, they are *content-based* in that they help users configure moderation choices based on the characteristics of the content they encounter on the platform. Examples of tools for personal content moderation include toggles, sliders, or scales for 'toxicity', 'sensitivity', or other attributes, as well as word filter tools for filtering out user-specified phrases (see Fig. 1, 2, 5). These differ from more common *account-based* personal moderation tools, such as being able to block or mute undesirable accounts individually or in bulk [30, 53, 82].

In the context of internet history, personal content moderation tools, with all the promises and perils they entail, have found their moment. Critics and scholars are increasingly calling for

mechanisms that move moderation decision-making away from centralized platforms and towards individual users to give them greater control over what they *do not* want to see on social media [26, 66, 74, 79]. Popular platforms like Facebook, Instagram, TikTok, and Twitter now offer word filter tools [52] and sensitivity settings [46] for users to configure over their news feeds and over 'Explore', 'For You', and 'Search' products. In addition, emerging platforms and third-party personal moderation tools like Gobo Social [10], Bodyguard [9], Block Party [82], and Bluesky [38] are letting users personalize their content moderation with even greater granularity. Social media founders and owners such as Mark Zuckerberg and Jack Dorsey have advocated for personal content moderation settings that tune exposure to controversial content, such as nudity, violence, and profanity [23, 119]. A recent representative survey of 983 adult internet U.S. users documented the public appetite for personal content moderation by showing that 52.4% of participants prefer relying on a personal moderation setting over the default platform-wide moderation to handle hate speech posts [55].

Despite growing interest in and adoption of personal content moderation tools, *empirical* insights about users' preferences for and attitudes towards such controls are scarce. Instead, prior research on content moderation tools has primarily focused on tools used by people in moderator roles in a community-wide setting, such as volunteer moderation teams [50, 75, 76, 92, 116]. This article aims to fill this critical gap and uncover the prominent user preferences, concerns, and design considerations regarding personal content moderation.

First, we sought to understand what considerations come into play when users have the ability to decide on personal moderation settings. A more restrictive configuration of settings could result in over-moderation, and a more lax configuration could lead to under-moderation. Thus, we ask:

> **RQ1:** *What considerations do users have when deciding whether to be more or less restrictive in their choices for personal content moderation settings?*

Next, we wanted to understand users' challenges when interacting with and configuring personal content moderation tools as realized in different commonly-deployed designs, including toggles, word filters, and sliders. We therefore ask:

> **RQ2:** *What issues do users encounter when trying to understand and control what happens when they interact with different interface designs for personal content moderation?*

Finally, since the enactment of any voluntary moderation tool targeted at platform users is inextricably linked to the questions of labor involved (e.g., see [24, 88, 95, 116]), we raise the following research question:

> **RQ3:** *How do users perceive the labor in configuring these tools? How do they ascribe the distribution of responsibility for this labor between different stakeholders?*

To gather deeper reflections and nuanced rationales for users' preferences, we conducted qualitative, semi-structured interviews with 24 social media users. We considered that participants may not have interacted with personal content moderation tools on social media platforms. Therefore, we developed a series of probes to prompt interviewees to consider different potential tool designs and elicit more informed opinions. To begin, we built a web application that simulates a social media feed of content along with a series of interactive controls that, when adjusted, re-configure the feed. We preloaded this web application with data from a research dataset of tweets labeled with different toxicity levels [64]. We implemented four types of controls based on personal moderation tools that have been deployed or proposed in the past (Fig. 3): a word filter, a toxicity toggle, an intensity slider, and a proportion slider. We describe these controls in detail in Sec. 3.

In the interviews, we asked participants to interact with all four control interfaces and inspect the resulting changes in the simulated feed while speaking aloud about their preferences. In addition,

we showed them screenshots of personal content moderation controls developed in industry, including the Bleep tool and one offered by Instagram. We used these probes to get more specific feedback on different design decisions. These probes also helped surface deeper reflections when we asked interviewees about their perspectives on whether and why they would (not) use such tools, moderation labor on the part of users, and platforms' responsibilities in the face of harmful content.

We found that while encountering offensive content on social media is a common experience, some prefer to just ignore such content, while others configure personal moderation settings, and still others get frustrated enough to quit social media altogether. Many interviewees were resistant to setting restrictive filters due to their fear of missing out on relevant posts and their desire to hear others out, even if the content might be offensive. Our analysis also raises critical areas for improvement in the current designs of personal moderation tools from the perspective of end users: increasing clarity in the definitions of various interface elements, incorporating environmental/cultural context, offering appropriate levels of granularity, and a wider leveraging of example content as a means to provide transparency and enable greater control. Our findings also highlight users' understanding of the cognitive labor involved in personal moderation and, related to it, their perspectives on how platforms and lawmakers share some responsibility.

We conclude by discussing the importance of addressing online harms while attending to users' desire to not overlook relevant content and how more context-aware personal content moderation tools can contribute to this goal. We highlight the value of clarifying the meaning of crucial interface elements and how doing so may necessitate an overhaul of current tools. We argue that designs that let users configure moderation exceptions for specified user groups or cultural contexts would increase the controllability of these tools. We emphasize that configuring these settings should be iterative for users, and incorporating user preferences inferred from other interactions, such as reporting and seeking user feedback, could further improve their utility. Finally, we examine platforms' roles and responsibilities in this space and how third-party developers and lawmakers can contribute to improving users' online experience.

2 RELATED WORK

2.1 Content Moderation

Content moderation is the organized governance of user-generated posts by information intermediaries and social media platforms to facilitate cooperation and prevent abuse on their sites [39, 88]. One of its main goals is to address online harms [56]. However, interpretations of online harm vary across cultures [57], communities [112], and individuals [51], making it challenging to deploy one-size-fits-all moderation solutions. We focus on content-based online harm that results from viewing undesirable content, such as hate speech or graphically violent images, on social media platforms. Prior HCI research has contributed to our understanding of the diversity of online harms [53, 90] and users' experiences of such harms [90]. We add to this conversation by examining how end users personally grapple with the questions of trading off viewing content that may be harmful to them against their fears of missing out on relevant content or desire to be open-minded.

In their early years, content on platforms like Facebook and YouTube was governed by relatively small review teams. Platform moderation rules and policies were also limited in scope [60]. As time passed and public pressure to remove offensive speech increased, platforms developed complex, sophisticated systems to aid their moderation functions (Table 1). Many platforms now have automated site-wide filters that remove, in an initial pass, blatantly inappropriate posts, such as spam or child sexual abuse materials (CSAM) [33]. Additionally, platforms deploy paid [88] or volunteer [76] moderators to review and regulate the remaining posts. Many social media platforms

	Account moderation			Content moderation		
	Actions	**Purview**	**Impact**	**Actions**	**Purview**	**Impact**
Platform moderation	ban or suspend users from the platform	all users on the platform	every user's view	remove content from the platform	all content on the platform	every user's view
Community moderation	ban users from a community	all users on the platform	every user's view	remove content from the community	content posted to the community	every user's view
Personal moderation	block a user from seeing one's content; mute a user from appearing in one's view	all users on the platform	one's own view; blocked user's view	remove content from one's view	content that appears in one's view	one's own view

Table 1. A characterization of different modes of content moderation. We have highlighted cells relevant to personal content moderation, this article's focus.

are also feed-based, incorporating algorithms that sort posts based on users' prior interactions [28]. Each social media platform has developed various ad-hoc systems to implement these processes [33, 100], yet each platform keeps the specifics of how it enacts its moderation decisions opaque [50]. We add to this literature by surfacing social media users' perspectives on platforms' moderation apparatuses and what role, if any, users should play. We also consider how a greater emphasis on personal moderation might reconfigure how platforms conduct moderation more broadly.

Besides top-down moderation, platforms offer several technical mechanisms for users to shape what they see. First, users primarily shape their feed by choosing to *follow* (or *unfollow*) or add as a *friend* any account from which they wish to see more (or less). Second, users can express their perception of any post by one-click mechanisms such as *like* or upvote/downvote. They can also report inappropriate posts using flagging tools, which trigger a post review by the site [17]. Third, some feed settings can more broadly shape the news feed, such as the settings to change the look and feel of the site, account deletion or deactivation, and language or region settings [44].

In addition to the above three mechanisms, users can also rely on *personal moderation tools*. These tools let users configure their preferences for the activity they want to *avoid*. Personal moderation tools fall into two categories: account-based and content-based. *Personal account moderation tools* let users mute or block *an account* to prohibit future interaction with it. Some third-party developers have built upon this functionality to enable mass blocking [30, 82] or peer-assisted blocking [71], and researchers have examined the utility and deficiencies of such tools [30, 53]. On the other hand, *personal content moderation tools* let users make moderation decisions *on each post* based on its content alone and regardless of its source. Personal content moderation tools include tools to mute specific keywords [52], remove[2] NSFW (not safe for work) content, and set up sensitivity controls (see Fig. 2 and 5 for examples).

In our work, we focus specifically on personal content moderation tools. We choose this focus because it has become increasingly common that the posts shown in users' feeds are not made by people they follow—an example is TikTok's 'For You' page, which is typically dominated by stranger accounts. The same can be said for users who use the search functionality on platforms such as Instagram or Pinterest to find content. In these scenarios, personal account moderation

[2]Note that in the context of personal moderation tools, post removals occur only for the configuring account. Others users can still see the removed content.

tools may not have as much impact on what users see due to the high proportion of novel accounts. This issue may also crop up in specific contexts, such as some gaming platforms where users frequently communicate with strangers [63]. As a result, personal controls that enable users to configure moderation based on content become more important. Indeed, in recent years, platforms have begun offering more personal content moderation tools [46], and industry leaders have called for more user controls for moderation [23, 119], leading to experimental efforts like Intel's Bleep (Fig. 1) and Jigsaw's Tune (Fig. 2(d)) tools.

As far as we know, this is the first paper that conceptually identifies personal content moderation as a distinct category and investigates user preferences for them. Prior research that focuses specifically on personal content moderation tools is scarce. One related work is Jhaver et al.'s research on need-finding and design exploration around word filter tools [52]. However, they focus on word filters used by content creators to delete comments on their content for all viewers, not users muting content in their personal feed. We note that specialized categories of users—such as advertisers, admins, and community moderators—have a broader range of moderation tools available to them, such as Automod [50] or Sentropy [41]. Prior research on designing such tools identifies some challenges relevant to our work, e.g., rule-based configuration producing false positives [13, 50]. However, our focus here is on moderation tools available to general users that do not rely on collaborative moderation teams and that only affect each user's own view.

2.2 User Control and Transparency in Interactive Recommender Systems

This section briefly reviews prior literature on interactive recommender systems as it offers valuable insights for designing personal moderation tools. Recommender systems are ubiquitously deployed today to solve the problem of information overload and to increase engagement [48, 58]. Over the past decades, extensive research has been conducted to develop and deploy algorithms that suggest relevant items to a user [3]. However, several challenges still need to be solved that prohibit recommender systems from realizing their full potential [42].

There is a growing awareness that the effectiveness of recommender systems goes beyond recommendation accuracy [11, 62, 86]. Research on integrating human values such as diversity, serendipity, and trust [58, 65, 96] into recommendations is gaining interest. As part of this push into user values, which may differ depending on the user, user controls like toggles and sliders have been proposed for interacting with recommender systems. This integration supports personalization, transparency, and controllability of the recommendation process [42]. These techniques form *interactive recommender systems* [42, 72], and the personal content moderation tools we study in this work offer similar controls. Advanced interactive recommender systems also incorporate contextual information (e.g., location, current activity, interest) to generate recommendations that tailor to the current needs of the user [4]. We examine the importance of incorporating relevant context in personal moderation tools from end users' perspectives.

In general, recommender and content moderation systems are similar in that they both act on the content one sees, where tuning a recommendation algorithm to reduce a specific type of content may have a de facto effect akin to moderating it out [34]. When done by platforms, this has often been described critically as "shadowbanning" [16]. However, a significant point of differentiation between the two is the goals behind them. Recommender systems aim to shape what content is shown to the user to make a better recommendation [48]. In contrast, moderation systems focus on removing content types that are likely to harm the user. As a result, users may have very different ideas about how personal controls for recommendation versus moderation should be designed. For instance, a user seeking to remove certain harmful content might desire more precise controls to be more confident they will not encounter that content compared to a recommendation control.

Indeed, in major platforms, recommendation and moderation often have different policies and separate teams dedicated to them.

Still, there are some aspects that users may find desirable in the design of both kinds of controls. Thus, we consider what we can learn from the interactive recommender literature. For instance, interactive recommender systems aim to provide *transparency* into the black-box nature of a recommendation system by explaining its inner logic to end users [94, 110]. Exposing the reasoning and data behind a recommendation may help increase users' confidence in that recommendation and improve their acceptance of the system [1, 43]. We examine how to provide transparency in the context of content-based personal moderation tools, where there is also an inner logic to the filter implementation.

Second, a related objective of interactive recommender systems is to offer *justification*, which refers to describing *why* the user gets a specific recommendation rather than *how* that recommendation was selected or how the system works [102, 110]. In some cases, justification may be preferred over transparency as descriptions of the underlying recommendation technique may be too complex or cannot be revealed to protect intellectual property [43, 102, 110]. We take inspiration from this literature by examining whether offering examples can help with understanding and if examples improve user acceptance of such tools.

The third key objective of interactive recommender systems is to improve *controllability*, or how much the system supports users in fine-tuning their configurations. Controllability can compensate for deficiencies in recommendation algorithms by incorporating user input and feedback to tailor recommendations to users' changing preferences [42]. Prior research has shown that in some application domains, users appreciate being more actively involved in the process and in control of their recommendations [22, 61, 81, 89, 105]. In the same vein, lower levels of user control can negatively influence the perceived quality of recommendations [40]. On the other hand, too many control interfaces can make user interfaces challenging to understand [5] and increase user cognitive load [58]. We examine users' perspectives on controllability in personal moderation tools by comparing interfaces that offer different granularities of control.

In summary, we take a user-centered approach in this work to examine how the types of personal moderation interfaces that commonly appear on mainstream social media platforms serve the preceding objectives of transparency, justification, and controllability.

2.3 Responsibilities of Platforms and Lawmakers

With the dramatic migration of online discourse to social media platforms in recent years, questions about how they structure social activity and the rights and obligations they have for the speech they facilitate are becoming increasingly important [18, 70, 80, 108, 111]. In the United States, Section 230 of the U.S. Communications Decency Act (CDA) protects intermediaries (including online platforms) from liability for illegal content sharing by users while also protecting them from liability when they do remove content that violates company policy [36]. This law reflects a fundamental reluctance to constrain speech inside the U.S. Most countries in the European Union and South America also do not hold platforms liable for their users' illegal posts as long as they comply with state requests to remove such posts [70].

In practice, however, nearly all platforms go beyond the legal requirements for removing inappropriate content, and user experience is shaped much more by platform policies than by legal restrictions. Critiques about these policies often mirror longstanding debates about the character of acceptable public discourse. Gillespie highlights examples of platform governance controversies that reflect society's broader public discourse concerns:

> *"which representations of sexuality are empowering and which are explicit, and according to whose judgment; what is newsworthy and what is gruesome, and who draws the line; how do we balance freedom of speech with the values of the community, with the safety of individuals, with the aspirations of art, and with the wants of commerce"* [32].

Gillespie argues that we should not leave the responsibility of resolving these fundamental tensions of social and public life to platforms alone but instead govern collectively as citizens [33]. Personal moderation tools empower end-users to specify their boundaries of acceptable conversation. In contrast, if we rely on platforms alone to serve acceptable public discourse for everyone, policies that enforce the most sanitized content, such as only what is appropriate for underage audiences or the content that Silicon Valley workers find acceptable, may prevail. Such policies may result in unnecessary censorship of otherwise contextually appropriate discourse. [117].

Many legal and media scholars have theoretically examined the dynamics of free expression online [8, 35, 67, 69]. However, there needs to be a more empirical understanding of how end users perceive their role in content moderation systems and where they might draw the line for themselves if given the option. Adding personal moderation complicates the role of platforms and policymakers in striking a balance between regulating inappropriate posts and protecting free speech, as platforms could leave up more borderline content and let users decide whether they want to see it. However, this raises critical questions about the physical and emotional labor involved in enacting moderation and the ethics of exploiting the unpaid work of end users to improve platform offerings [24, 50, 88, 95, 116]. It also raises ethical questions about what speech users should get a say in versus what speech platforms should be compelled to take down for all users. We leverage our participants' examination of a sample of personal moderation interfaces to induce them to reflect on the questions of labor and the responsibilities of different stakeholders in this space.

3 METHOD

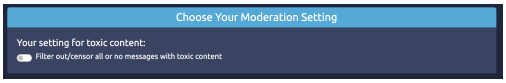
Our university's IRB[3] approved this study. Our study involved 24 semi-structured interviews with active social media users. During the interviews, we used a technology probe [45]—in this case, an interactive, toy social media feed that simulated how four personal moderation interfaces would work—along with additional screenshots of existing tools to prompt each participant in order to answer RQ2. Every participant interacted with the four interfaces. Additionally, we asked participants questions about the contexts in which they could use these tools, the tradeoffs they consider in using them, and their perspectives on the labor involved to answer RQ1 and RQ3. We now describe the details of our study design.

3.1 Simulating Personal Content Moderation Interfaces

We began by systematically observing personal content moderation tools and interfaces on the most popular social network platforms available in English.[4] These platforms included Facebook, Twitter, Instagram, and TikTok, among others. We created a new account or used our existing account on each site and looked into its settings page to observe the options available for user-enacted moderation. We also examined the options available through third-party moderation tools like Gobo Social [10], Bodyguard [9], and Intel's Bleep tool. Given our focus on personal *content* moderation tools, we did not consider account-based settings such as muting and blocking accounts. We focused only on tools available to end users engaged in consuming content rather than on community moderator tools, which often deploy more specialized automated moderation tools [50],

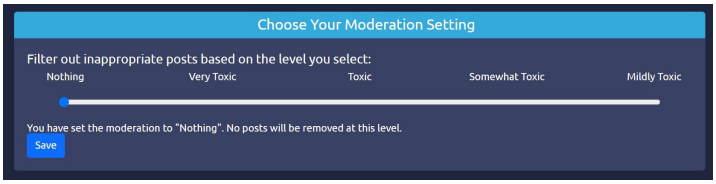
[3]We will disclose the University name after the peer review process is completed.

[4]We focused on the English language platforms with at least 100 million active users available on https://en.wikipedia.org/wiki/List_of_social_platforms_with_at_least_100_million_active_users

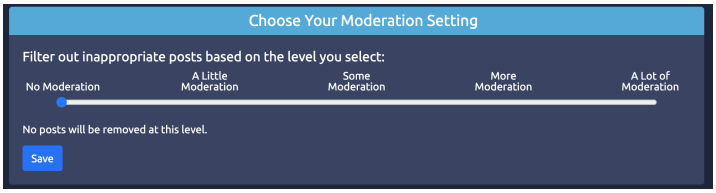


(a) Binary Toggle

(b) Word Filter



(c) Intensity Slider



(d) Proportion Slider

Fig. 3. Implementations of four moderation interfaces used in our study. (a) Binary toxicity filters. (b) Word filters. (c) Intensity sliders. and (d) Proportion sliders. These interfaces were inspired by personal moderation tools commonly available on popular social media platforms.

or content creator tools that moderate comments on their content [52]. Through these observations, we concluded that the commonly deployed moderation interfaces included:

(1) *Toggles.* These tools offer a 'yes/no' choice so users can allow or avoid seeing potentially undesirable content. For example, Twitter offers on its settings page a checkbox for "Display media that may contain sensitive content", and TikTok has a 'Restricted mode' that may be turned on to restrict 'inappropriate' videos. Tumblr and Bluesky have toggles for different categories of content, as well as a 'blur' or 'warn' option in addition to showing or hiding.
(2) *Word filters.* These tools let users configure a list of phrases; posts containing any of those phrases are automatically removed or moved to a separate folder for human review. Word filters are commonly available on most popular platforms for muting content on one's feed, including Twitter, TikTok, and Instagram.
(3) *Sliders.* These tools let users specify on a 3–5 point scale their desired thresholds for a given concept over the posts they see. Instagram has a 3-point scale for 'sensitivity' but presents options in a list with radio buttons (Fig. 5). Intel's Bleep has 4-point sliders for a number of categories. Most versions of sliders have levels according to proportion of moderation, but Gobo Social's levels are according to the intensity of the concept.

We created a website to simulate a user's experience viewing a news feed of comments on a social media platform much like Twitter. Based on our observations of popular moderation interfaces

described above, we also built four moderation interfaces on this website that let users change their personal moderation settings for each interface and observe a changed news feed. These interfaces included a toggle for 'toxicity', a word filter, a 5-level slider for 'toxicity' that has levels labeled according to intensity or degree of toxicity, and a 5-level slider for 'toxicity' that has levels labeled according to proportion or amount of moderation; see Fig. 3 (a-d). We built and used these interfaces as interview probes for eliciting user opinions on the design space of personal content moderation tools. The intensity and proportion sliders were set to the default option of "no moderation."

We focused on these interfaces because they afforded different levels of granularity, transparency, and control in their design—themes we wanted to explore in our user study. The set of interfaces we developed is not an exhaustive representation of the currently available moderation tools. Further, some available tools, e.g., Intel's Bleep or Tumblr's labels, have a more detailed granularity of moderation settings. We selected and built these interfaces to demonstrate to interviewees ways in which the design of personal content moderation tools could vary. We also sought participants' views on additional tools like Intel's Bleep and Instagram's sensitivity filters to aid this design exploration further.

We implemented our interfaces for only text comments as labeled datasets for toxicity were more readily available for textual rather than multimodal content. Still, text-only comments form the bulk of content on many platforms. We also asked our participants' views on multimodal personal content moderation tools offered by Intel and Instagram.

3.1.1 Comment Curation. To curate comments for our simulation website, we began with a labeled dataset of 107,620 comments from Twitter, Reddit, and 4chan [64]. We filtered for only the Twitter comments because comments from other platforms were more difficult to understand out of context. Moreover, these platforms' thread structure differs from the Twitter-like feed we simulated in our study. We used the following attributes available for each comment in this labeled dataset: (1) its text; (2) whether the comment was profane, a threat, an identity attack, an insult, or sexual harassment; and (3) five independent ratings for each comment on a five-point scale from 0-4 ranging from "Not at all toxic" to "Extremely toxic."

Next, we filtered only those comments for which raters in the dataset had a consensus on the toxicity levels. To achieve this, we excluded all comments where the difference between the highest and lowest rating was greater than one point. Next, we calculated the average of the five toxicity ratings for each comment and bucketed all comments into five windows of equidistant toxicity averages. We call these buckets B1, B2, ..., B5, with B1 representing the bucket with the lowest average toxicity and B5 the highest. From each bucket B2–B5, we selected a random sample of 20 comments.

All coauthors then manually reviewed these 80 comments and selected 5 comments from each bucket that were comprehensible without additional context. We selected comments that, taken together, represented topical diversity (e.g., cooking, food, sports, politics, celebrity news) and diversity of inappropriate behaviors (e.g., profanity, threats, identity attacks, insults, sexual harassment). We achieved this through mutual discussions and clarifications. We also selected an additional random sample of 30 comments from bucket B1.

3.1.2 Interface Implementation. We simulated our news feed to show 30 comments by default: 5 comments each from buckets B2–B5 and 10 from B1, randomly shuffled together. We implemented each moderation interface as follows:

(1) *Binary toxicity toggle.* By default, the toggle setting was turned off. When the toggle setting was turned on, we removed comments from buckets B2-B5 and showed only the 30 comments from bucket B1.

(2) *Word filters.* We excluded all comments matching the keywords configured by the user in their word filters. We replaced the excluded comments with additional bucket B1 comments to maintain the total number of comments at 30.
(3) *Intensity slider.* This slider removed progressively more toxic comments as the slider level moved from 'Mildly Toxic' to 'Very Toxic'. For example, when users moved the slider level from 'Nothing' (default level) to 'Very Toxic', we removed five comments from bucket B5. We replaced them with 5 randomly selected bucket B1 comments to maintain the total number of comments at 30.
(4) *Proportion slider.* This slider's implementation pre-classified each comment as non-toxic or toxic, depending on whether the comment was in bucket B1 or not, respectively. When moved to its immediate right, each slider level removed 5 randomly selected toxic comments from the feed and replaced them with five bucket B1 comments.

Our website lets users choose one of these four moderation interfaces at a time. Once a new interface is selected, the settings for all other interfaces are programmed to reset to their default levels. For example, if a user set up a few keywords in word filters and then used the intensity slider interface, our implementation would not filter out the keywords configured in word filters. This setup resulted in the absence of interaction effects, which made it easier for users to understand the operations of each new interface they used.

3.2 Participant Recruitment

We recruited interview participants for our study by advertising on social media platforms such as Facebook, Twitter, and Reddit. Our calls for participation asked candidates to submit an online form that helped us obtain relevant information about the candidates. This form included questions about whether the participants used social media daily, which social media platforms they used, whether they encountered toxic content on social media, and their demographic information. We also included an open-ended question: "What is your perspective on how social media platforms should deal with toxic content?"

Analyzing the responses to this form, we selected candidates who seemed to have some familiarity with or be more likely to benefit from personal moderation tools, given their experiences, since our focus was on how best to get insights about their use and potential improvements. We examined the depth and clarity of users' responses to our open-ended question and their previous encounters with toxic content to guide our participant selection. We also oversampled individuals from marginalized groups, such as Black and LGBT users, since such users are more likely to experience online harm [91] and, therefore, could benefit from more advanced moderation tools. Further, we ensured that our interview sample represented diverse ages, occupations, and countries. Table 2 shows a list of the participants.

3.3 Data Collection

Our interviews lasted an average of 81.65 minutes (sd = 12.16 minutes, range: 60 - 94 minutes) and were conducted over Zoom, recorded, and transcribed. Participants completed an informed consent form before proceeding with the interview. We explicitly informed participants at the start that we would record the interview to enable analysis and answer our research questions.

To understand the role of personal moderation tools in situated contexts, we first asked participants about their general social media usage during the interview. Next, we asked whether they encountered offensive content on their news feed, what action they took in such cases, and what happened as a result. Next, we briefly reviewed the goals of our session and requested a screen-share before asking them to use our website. We explained that our purpose was not to

#	Age	Gender	Occupation	Country
P1	27	Female	Grad student	USA
P2	24	Female	Grad student	USA
P3	23	Female	Grad student	USA
P4	31	Male	Grad student	USA
P5	34	Male	Journalist	UAE
P6	26	Male	Grad student	USA
P7	21	Female	Student	USA
P8	26	Female	Grad student	USA
P9	25	Male	Grad student	USA
P10	40	Female	Stand-up Comedian	India
P11	40	Male	Angel Investor	India
P12	56	Male	Editor	India
P13	23	Female	Engineer	USA
P14	44	Male	Applications Manager	The Netherlands
P15	48	Male	Documentation Specialist	Belgium
P16	34	Male	Civil Engineer	UK
P17	26	Male	Journalist	UK
P18	29	Female	Physical Therapist	USA
P19	39	Male	Psychologist	Canada
P20	33	Male	Director of Content & Marketing	USA
P21	23	Male	Student	Australia
P22	36	Trans male	Mechanical Engineer	UK
P23	33	Male	Software Engineer	USA
P24	24	Female	Student	USA

Table 2. Demographic details of participants with whom we conducted semi-structured interviews for this study.

create a new social network site but to use our simulation of different interfaces as probes to elicit their perceptions of using different personal content moderation tools. We requested that participants "think-aloud" [68, 109] as they tried different interfaces, expressing their likes, dislikes, and confusion and describing additional features they would like to have.

During these sessions, participants frequently toggled between news feeds and the settings page under each interface to observe how changes to settings altered their feeds. We asked participants to reflect on how they would configure different interfaces if they used them to moderate their news feeds. We clarified to the participants that their configurations would affect only the content on their news feed, not other users' feeds. In these interviews, we avoided explaining what the different interfaces did without solicitation but explained their purpose and functions when asked. This strategy let us understand the differences between participants' understanding and the actual function of interfaces when such differences existed.

For our intensity slider, we created an additional interface that, for each level, showed examples of comments that would be removed at that level (Fig. 4). We asked participants to comment on whether these examples influenced their perceptions of this interface.

In addition to our interface implementations, we showed participants screenshots of categorical toxicity sliders implemented by Intel (Fig. 1) and asked them for their thoughts on it. We also sought feedback on the sensitive content control settings implemented by Instagram (Fig. 5) since this is a deployed variation of the sliders we implemented on our website. We encouraged participants to compare and contrast different interfaces based on their efficacy, explainability, and flexibility.

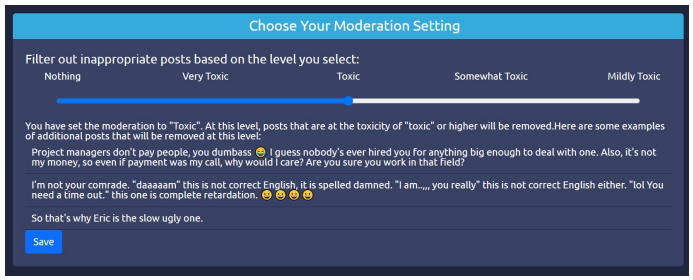


Fig. 4. Intensity slider with examples of comments that would be additionally removed when toggling to each level.

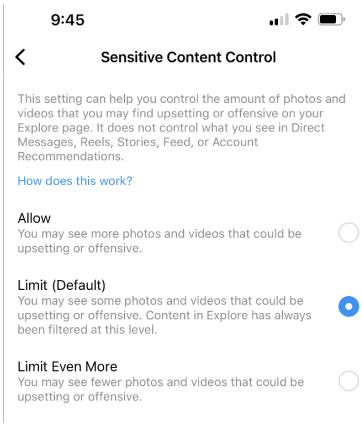


Fig. 5. Screenshot of sensitive content control settings implemented by Instagram.

Immediately following this hands-on session, where participants configured different moderation interfaces, we asked them about their views on the labor involved in setting up such configurations and the role that platforms and policymakers should play in ensuring appropriate content curation. Finally, we asked participants for their demographic information. After the interview, we compensated each participant with a $20 Amazon gift card (or an equivalent amount for foreign countries).

3.4 Analysis

We began our analysis by reading the interview transcripts and familiarizing ourselves with the data. Next, we applied interpretive qualitative analysis to all interview transcripts [77]. We uploaded all transcripts to a dedicated project in Dedoose,[5] a cross-platform app for analyzing qualitative research. Then, we conducted "open coding" [15] on a line-by-line basis so that codes stayed close to the data. Through this coding, we categorized segments and created codes that summarized and accounted for each piece of data in concise terms. Examples of codes in this stage included "Desiring attack notifications" and "Seeing offensive content."

Next, we performed multiple iterative rounds of coding and memo writing. During this process, we continually compared our codes to their associated data. To do so, we paid close attention to the

[5]https://www.dedoose.com

dilemmas and tradeoffs that emerged in shaping user preferences, e.g., desiring more control over moderation tools versus exerting more configuration labor. Our memo writing helped us remain open to emerging themes and deepen our reflections. The authors discussed the codes, memos, and emerging concepts throughout our analysis each week. We contacted some participants to clarify their responses during the analysis stage further.

After the first round of coding, which stayed close to the text, our next round was at a higher level. It resulted in codes such as "Appreciating the ability to fine-tune" and "Labor involved in setting category-based filters." Through the following rounds, we combined and distilled our codes into key themes we next present as our findings.

3.5 Positionality Statement

As researchers working in the sensitive space of understanding online harm and imagining interventions that may help address it, we briefly reflect on our position on this topic. All the authors of this paper feel deeply concerned that online harm is a persistent social problem that disproportionately affects marginalized and vulnerable people [25]. Some authors also come from marginalized groups and are survivors of harm. Our prior research on online harassment has helped us understand the inherent limitations of traditional, platform-wide moderation mechanisms; this, in turn, has shaped our desire to look for alternative approaches. At the same time, we are also concerned with threats to free speech that overly restrictive and globally enacted moderation solutions pose. In examining and improving the design of personal content moderation tools, we seek to empower users, especially from marginalized groups, to participate in online discourse freely, safely, and on their terms.

4 FINDINGS

We begin by describing in Sec. 4.1 how participants react to experiencing online harms differently, emphasizing the trade-off they face between mitigating harms and missing out on valuable content (RQ 1). Sec. 4.2 details the challenges in using personal content moderation tools that participants identified in our interviews (RQ 2). This subsection offers insights from participants' direct engagement with interview probes, including our toy website and screenshots of existing personal content moderation tools. Finally, Sec. 4.3 and 4.4 present participants' perspectives on the labor involved in moderation configurations and how platforms and lawmakers could better serve them, respectively (RQ 3).

4.1 Reacting to Online Harms and Concerns about Missing Out on Valuable Content

Our participants reported a wide range of harms they encountered on social media sites. They experienced harmful content not just on their news feed, but also in comments to their posts, direct messages, and profile pictures. They expressed varying responses to such harms and discussed the intention to stay online despite them. Participants also reflected on the trade-offs between mitigating harmful content and missing out on valuable content. We now detail these reflections.

4.1.1 Responding to Online Harm. Participants' responses to experiencing online harm vary depending on their interpretations of online interactions and the tools made available on specific sites. These responses included using the available moderation mechanisms, ignoring online harms, and quitting social media.

Our participants showed sufficient literacy of personal moderation affordances available on different platforms. Their mental models of what mechanisms such as blocking, muting, and reporting accomplish were aligned with the understanding of authors, who are content moderation experts. Many participants deployed mechanisms including reporting or muting offensive posts

and reporting, blocking, unfollowing, or muting perpetrator accounts. Participants noted that they needed to engage in such actions to protect their mental health. For example, P20 began "hiding" content he deemed not good for his mental health to avoid repeated exposure. However, in line with prior research [53, 63], many participants noted unsatisfactory experiences with such actions, especially with reporting inappropriate content.

Some participants preferred not to use any of the moderation mechanisms offered by the platforms. For example, P7, P21, and P22 preferred to ignore and disengage with harmful content. P21 described his reasoning for this behavior as follows:

> *"If one person's being racist, there is probably a lot of other people also doing that, and there's not much of a point for me sifting through every single comment and reporting it. Usually when I see that kind of thing, I just call it [a day] and go do something else."* – P21

On the other end, several participants stopped using social media sites altogether due to experiencing harmful content. Both P18 and P22 stopped using Twitter after seeing too much harmful content. P22 used to rely on social media sites to get his news, but due to the negativity on these sites, he turned to view news outside social media. These reactions suggest that efficient, user-friendly personal content moderation tools could benefit some users by letting them address online harms and continue engaging in online discourse.

4.1.2 Concerns around Personally Moderated Content. Participants worried about what they might miss if they used any of the interfaces they sampled to censor content for themselves. For example, P24 felt that her social media feeds were already "too much of an echo chamber" and that using content moderation tools would "over-censor" what she sees. P14 reflected on his "fear of missing out" (FOMO), arguing that he worried more about missing important content than encountering toxic posts. Many participants did not want to shelter themselves on an online platform from the harsh realities of offline life. P24 appreciated when Instagram flagged sensitive content, like images or videos of severely underweight people, but did not stop users from seeing it. P11 reported that his social network news feeds once served up a video of a beheading. He was troubled by it but would not want even such gory content to be filtered out:

> *"If you go to a party and somebody is talking about that beheading that has happened, I will get exposed to it. It's not like I can cut out those conversations offline." - P11*

Some participants also hesitated to suppress inappropriate posts because they desired to remain informed and take appropriate actions in response to such posts. For example, P10, who received rape threats, emphasized the importance of not suppressing threatening posts that could serve as warning signs for impending offline harm. P3 noted that if any account she *followed* began posting misogynistic content, she would want to know about it and take actions such as *unfollowing* that account or reconsidering how she related to that person. P1 did not mind seeing emotionally volatile posts, even if they were offensive, because she *"wanted to know the reality, like what people are angry about and if there is anything I can do to help them."* Three participants felt strongly about their responsibility to keep online spaces safer for everyone by observing and then flagging inappropriate content. Further, many of our participants were apprehensive about the ability of machine learning-based moderation systems to effectively moderate and therefore preferred the option of minimal or no filtering.

4.1.3 Inclination to Remain Open-minded. Participants also reflected on the balance between avoiding online harms and remaining open-minded. Everyone agreed that egregiously inappropriate content, such as calls to violence or malicious disinformation, went beyond the limits of free speech and should be taken down for everyone. P11 argued that free speech rights came second to local jurisdiction and that any posts that violated local laws should be removed. Beyond that, many

interviewees felt that platforms should respect the users' First Amendment rights and allow everyone to speak freely. As we will discuss further in Section 5.1, this is a misunderstanding of First Amendment rights since platforms are not arms of the government but private entities that are free to censor anyone they like [31, 60].

Participants described their desire to remain open-minded when consuming social media content in the context of personal content moderation tools. Some argued that personal content moderation controls should not exist because everyone should be more open-minded. Overall, participants recognized the difference between their actions to hide a post via personal content moderation vs. a platform's removal of that post for everyone. For instance, P14 did not consider his hiding a post as censorship since others could still see that post.

4.2 Understanding and Engaging with Personal Content Moderation Tools

Our participants were keen to interact with the various personal content moderation tools we presented to them through our toy website. They also expressed how these tools could address their moderation challenges and improve their social media experience. Many participants were especially excited about the more granular moderation available through Intel's Bleep interface. Since prior research [55] has documented the need for such tools, we focus here on the nuances of where users see room for improvements in their design.

Our participants reported encountering challenges when engaging with personal content moderation settings. First, they regretted that the terms used to ground various elements of the personal moderation interfaces were often not precisely defined. Second, they felt that the systems' failure to account for critical environmental contexts and offer appropriate levels of granularity hindered their ability to configure aspects of personal content moderation to their liking. Finally, participants reflected on the systems' use of examples to offer explanations and instill transparency. We detail these findings below.

4.2.1 Defining the Terms and Criteria. We found that participants needed to build a mental model around how the platform carries out personal content moderation before they could more comfortably and deeply engage with moderation settings. While the mental model of how word filters operated—comments containing the configured keywords would be automatically blocked—was generally clear to users, they had more trouble forming mental pictures for the slider and category-based settings. For these settings, platforms present the moderation system as an automated tool that categorizes content into high-level groups such as *"hateful speech"* or *"sensitive content."* These groups are described by the displayed label and some short text descriptions to direct configuration changes (e.g., "Feel free to fine tune" used by Intel, see Fig. 1). In such cases, users often desired greater transparency and clarity about the criteria that define these categories.

Ambiguous Definitions. When personal content moderation settings were presented as a categorization of content, either by type (e.g., "sexism," "racism") or along a scale (e.g., "very toxic," "somewhat toxic," or "A little moderation", "Some moderation"), participants often wanted to dig deeper into the definitions of these terms. They found the term *toxicity*, used in our slider-based interfaces, to be especially ambiguous and subjective. For example, P16 commented, *"I can see it working in a black or white manner, but the idea of toxicity—that's very subjective!"* P7 noted that due to the variability of community norms and topics of interest, *"one thing that can incite hate among one community [can be] kind of benign to another,"* so it can be challenging to understand the platform's criteria. P22 also noted that different people use language differently, so it may be undesirable if linguistic characteristics, such as the presence of swear words, were used as criteria for toxicity.

Participants also observed this perceived ambiguity of meanings in other interface elements. When examining the Intel category-based filters, both P2 and P7 could not understand how the system defined category groups like "name-calling" despite their brief text descriptions in the interface since they found them imprecise. Participants noted that depending on the system's definition and criteria of such terms, they might not consider some instances of name-calling as harmful or offensive, which would guide their configuration actions. However, the lack of clarity in descriptions of relevant terms interfered with their sensemaking.

Besides the confusion around the meanings of terms like "toxicity" and "name-calling," participants also reported challenges in understanding what distinctions existed for the levels offered on slider scales ("What is in each level?"). For example, P22 expressed confusion around the classification process for levels on a slider control. It was important for participants to understand what each level meant in more concrete terms than the interface presented to determine whether this type of control was relevant to their moderation goals. Responding to a question about whether there was any additional information that could help him select a level on a slider interface, P22 responded:

> *"I mean, kind of, describe how it selects, how it is actually working! [Pointing to a comment] Because thinking about this one—they direct insulted someone. Well, that's pretty rude! [Changing settings and pointing to another comment on the changed news feed] But yeah, that one is calling someone a dumbass as well, so I do not know why this one is seen as worse than that one."* - P22

Evolving Definitions. Participants also pointed out that meanings of words and criteria used to define categories can evolve, so a static definition (such as *toxicity*) may need to be updated. When discussing the definition of *aggression* in one of the moderation interfaces we showed, P19 commented:

> *"You know, something like aggression, you can define aggression on an algorithm. But ableism, you know, [and] LGBTQ discrimination—those are constantly evolving concepts and ideas based on social parameters. And so, having a scale...it will be obsolete the moment it is rolled out. And it will be based on ideology, rather than actual concepts."* - P19

Lack of Transparency about Relevant Moderation Criteria. Participants were curious about the factors that moderation systems incorporate in their decision-making and how they resolve conflicting or competing preferences. For example, P5 wanted to know what is included in the input to the algorithms that determine toxicity, asking: "*How [is] extreme toxicity [determined]? Is this one classified depending on a word, or an expression, or the entire tweet itself? The sentiment of the tweet itself?*" Three participants assumed toxicity was determined based on the language used and described instances where it is inappropriate to operationalize toxicity based on the presence of certain words. For example, P19 felt, *"What makes something toxic isn't so much the word choices, but rather the emotion behind which drives the word choice."* Regarding interactions between overlapping criteria, as in the Intel moderation interface, P3 desired more clarity in scenarios where two topical filters may be in effect simultaneously:

> "*I think there's overlap in some of these categories. And so if I don't mind seeing name calling generally, but maybe I don't want to see anything having to do with misogyny, but there's going to be some overlap between those two categories. And then do I not see stuff? Or do I always see stuff? That is confusing to me. How are you to sort out aggression versus name-calling?*" - P3

4.2.2 Failure to Account for Environmental Context. Participants noted that many personal moderation systems did not or could not account for relevant context when moderating content, yet their desired moderation outcomes were often context-dependent.

Rigidity in Word Filters. In the preceding sections, we noted that participants found word filters easy to form a mental model around. However, their effects on news feeds were often undesirable in practice due to their inability to account for context. P21 gave an example of how this fluidity of word meanings could be a challenge when moderating news feeds in different cultural communities:

> *"I think a lot of these words that have been used depend on the context, like 'c*nt.' So, I'm in Australia, we say that all the time and usually it's not a bad thing."* - P21

This lack of ability to account for environmental context can affect not just neighborhood communities but also far-flung communities defined by shared subcultures or social concerns.

> *"You can say the exact same sentence but in one context, it is body shaming and in another context it is a compliment for somebody."* - P15

Accommodating Use of Reclaimed Slurs. Concerns about the context also arose for minority communities who reclaimed slurs to fight against historical oppression. For example, some participants speculated that the "n-word toggle" from Intel's category-based filters could be problematic for African-American communities that do not object to its use in in-group conversations. While they may wish to apply such a filter to content outside the community, an across-the-board filter that does not account for the poster's identity would also censor well-meaning conversations inside their community. Word filters enact a similarly crude moderation, where the only criterion for comment removal is the presence of any configured keyword regardless of other factors.

Insufficient incorporation of relevant context was also cited as a shortcoming of slider-based tools. P2 gave two examples of how inappropriate design of moderation tools can harm rather than help specific communities when commenting on the *body shaming* and *name-calling* categories of the Intel slider:

> *"The fat community is really moving forward with fat empowerment, and you don't want to shut down their posts because they're talking about fatness in positive manner."* - P2

> *"A straight person can call a gay person a twink and that's like derogatory. But like if a gay person calls another gay person that, that might not be derogatory. One of my best friends is trans and they'll use that term all the time when we're referring to their friends, that's just like how they talk about their friends."* - P2

In summary, when user moderation needs depend on environmental context, engaging with personal moderation systems that cannot account for such context or do not allow specifying context as part of the setting can be frustrating.

4.2.3 Granularity of Control. Another issue participants noted was that the granularity of control offered by various personal content moderation tools was often inappropriate. This problem was reflected not only in a lack of higher granularity but also through systems offering control over aspects that did not match the users' moderation goals.

Slider-based Controls. Some participants desired a higher granularity of control when interacting with personal moderation tools. After using a simple toggle to control toxicity moderation, P4 and P17 felt that the moderation seemed too extreme when turned on, leaving the content "way too overly positive." Most participants appreciated the increased degrees of freedom that sliders and category-based controls offered compared to a binary setting. For example, P2 and P11 appreciated the ability to adjust moderation levels across specific categories using Intel filters, noting that these sliders accommodated users with different levels of acceptance for each category. P21 called sliders a good solution "because you can kind of fine-tune it" but wished for even greater granularity, suggesting using a continuous, rather than a discrete, slider.

However, optimal granularity is not achieved simply by increasing the degrees of freedom in all cases. Some participants noted that adequate visibility into the behavior of moderation tools was a prerequisite to determining how much control was practicably achievable or even desirable. When differences between settings were not easily interpretable, participants saw little benefit in extra granularity. For example, when asked about the number of levels appropriate for the category slider, one participant noted:

> *"When I was looking at the five-level slider, there wasn't a huge difference between 1 and 2, and there wasn't a huge difference between 4 and 5. I would think you might be able to make that just a three level slider and not have a huge difference."* - P22

P1, P5, and P6 reported similar concerns for sliders with examples, noting that examples alone often did not provide enough information to infer what was being controlled at the different levels of the sliders. How we designed our toy platform can possibly partially explain our participants' inference of the lack of differences between adjacent levels. However, these responses suggest an essential guideline for any platform to consider when building slider-based moderation tools—they should strike an optimal balance between providing the right granularity and adequate transparency into the differences between various levels created at that granularity. P7 commented on the gap between system designers' tendency to simplify moderation settings into levels and the challenge of providing more in-depth, but also complex, control over personal moderation:

> "*It is sort of a web designer's idea of user-friendly in the sense of streamlining everything and putting it on the back-end, but I do not think that is necessarily the same as actually giving users, the client, the levels of choice that they might want, you know? I think, actually, in some ways, a more complex system that doesn't use sort of vague terms like toxicity levels will probably be better in some ways.*" - P7

Word Filters. Issues with granularity also arise with word filters, where including or excluding a single word is a fixed binary choice. For example, when word filters are used to moderate swear words, they can present a granularity of control that is too high, exposing users to details that they may not want to face. P9 noted that it can be hard to "pre-specify" what type of content he might be uncomfortable with: *"You know, you might just not be aware of what you do not want, what's going to bother you."* Even when participants had specific moderation goals in mind, they lamented that it can be *"a lot of work to think of every possible word that might be offensive"* (P24) and that *"it would be nice to get assistance on covering variants like slight changes in spelling"* (P12). P22 additionally noted that it would not be challenging for bad actors to get around word filters unless the filters had sufficient coverage:

> *"I can invent 20, 30 ways to say that word that you just blocked. [People] can write really, really ugly things avoiding any of the words."* - P22

Configuring personal moderation at a word-level granularity also forced users to confront harmful content uncomfortably. For example, P24 noted that *"I feel, like, very uncomfortable writing...the n-word."*

4.2.4 Leveraging Examples to Instill Transparency and Control. Throughout our interviews, we found the prominent use of example comments to make sense of and control the process of configuring personal moderation.

Using Examples to Align Mental Models. Many participants found examples to be a good way to build their mental models about moderation controls, especially when text explanations of relevant terms were missing. For example, P3 and P16 found it challenging to understand the specific meanings of various toxicity levels in intensity filters. However, examples helped them understand the types of comments that each level excludes. P6 additionally expressed that

examples not only helped understand moderation tools but could ease their taking of configuration actions—*"to understand, okay, this is what I have to do!"*

In addition to providing the necessary grounding to contextualize their mental model, participants noted that examples offered an excellent opportunity to identify when their mental model mismatched the actual behaviors of the moderation system. Some participants reflected on the potential problems with using real examples of offensive or toxic content. For example, P7 noted:

> *"It might be a little bit more helpful to show the criteria instead of the examples that would be filtered out because with examples, you have to see the things that you didn't want to see."*

Using Examples as a Form of Control. Examples provided a rich amount of information about the behavior of the moderation tools. As a result, many participants expressed a desire to use examples not just as a passive tool to provide transparency but as a direct means of control. For example, P13 foresaw the possibility of using manually written examples to define moderation goals directly or through training models based on labeling existing examples.

> "*One [way to use examples] is to have me type out an example of something, which I think would be kind of an unpleasant activity and also difficult to imagine right off the top of my head ... The other thing could be, as I scrolled through my timeline, as I encountered pieces of toxic content ... I could put it in these [undesirable] categories.*" - P13

Using Examples as Feedback for Control. In addition to understanding or controlling personal moderation, users relied on real examples to get feedback on their personal moderation settings. In this case, they utilized changes to their feed as examples rather than the examples presented by the system. We observed several participants switching between settings and news feed pages to verify that their configuration of the personal moderation tool achieved their desired goals.

> *"I think, to me, what's really interesting is that this post gets filtered out immediately when you change the moderation setting. I don't understand why it's getting filtered out. So I think... Let me fiddle with it some more."* - P2

Indeed, applying and testing out interactive control over moderation became a way for users to form more accurate mental models over personal moderation systems iteratively. However, this kind of feedback can be labor intensive:

> *"The more effort I invest into something, like the more time I spend, for example, in curating my Twitter feed, the more irritated I get when I get things on my feed that I didn't want."* - P2

We discuss these labor aspects in the next section.

4.3 Labor Involved in Configuring Moderation Systems

Many participants showed a keen awareness of the extent of labor involved in setting up and maintaining moderation configurations. P2 and P3 described this as a trade-off between the extent of agency they have over curating their social media feeds and the labor they must exert to configure moderation filters. Participants wanted to see improvements to their news feeds in proportion to the effort and time involved in setting up the corresponding configuration. For example, some participants felt frustrated with having the option of multiple levels in the sliders in our study because they kept seeing what they deemed inappropriate comments despite moving between different levels.

A few participants saw setting up moderation configurations as an additional burden they were unwilling to take on. For example, P23 said:

> *"I think I am a typical user - I'm probably using my phone at a time I'm really tired, I come back from work—I don't really care, and so if you ask me to customize to that level, my laziness would come out and I'd be like 'Nope I don't want to click so much.' "* - P23

Similarly, P18 preferred category-based filters over word filters because she used social media to "decompress" during her lunch breaks and did not have time to configure the inappropriate keywords worth blocking in word filters.

> *"If I have to go on there, and I feel like I'm designing or customizing or engineering and I'm doing this for free... no I don't want to do this, this is supposed to be fun!" - P18*

P21 felt that configuring topical category-based filters was onerous. He instead suggested a design with a main toxicity slider, where users could additionally search for and configure specific topic-based filters. P18 argued that configuring any of the interfaces we showed him would require his full attention; he found these configurations more demanding than other moderation actions, such as flagging a post or blocking a user. However, he also felt that such configurations would rarely need updating once set up.

Participants also considered the monetary value of labor involved in setting moderation configurations. P2 and P3 pointed out that in enacting moderation configurations, users produce valuable training data for the complex algorithmic systems platforms use and should therefore receive some compensation.

Some participants wished platforms would design ways to reduce the labor involved in moderation configurations. For example, P3 wanted word filters to have the ability to quickly configure entire groups of pre-configured keywords representing offensive categories such as sexism or homophobia. She also wanted to see which keywords her friends added to get ideas about the keywords to configure for her account. P1 wanted slider interfaces to indicate how the news feed changes in response to settings changes to reduce the work of going back and forth between the tool and the feed to observe what has changed.

4.4 Distribution of Responsibility

As a follow-up to our participants expressing their views about the different personal moderation tools, we asked them about their perspectives on the obligation of configuring such tools. Specifically, we asked for their opinions on how the responsibility for appropriate content curation should be distributed between the platform and the end users.

In response to this line of inquiry, most participants argued that platforms should, at a minimum, remove the most egregious content posted on the site. For example, P12 and P21 noted that many user groups, such as children or technically illiterate users, may be unable to set up their personal moderation preferences. Therefore, platforms are responsible for ensuring appropriate content delivery by default. P15 and P19 felt that given the secretive nature of content curation that platforms use, they have a social and moral obligation to remove inappropriate comments. P8 said platforms should not allow social unrest, like the January 6 Capitol attack, to be instigated on their site because it violates their policies.

> *"They should make sure that people stick to how Instagram and other sites originally intended users to share content. But like, if they're using it to incite people, to rile them up purposefully, or to just like using the platform in ways that are harmful, that is going to cause real harm. That becomes the responsibility of the platform. Because these people are like, you know, not using what they built properly."* - P8

In the same vein, P2 argued that platforms should consider the ongoing misinformation and hate speech trends and ensure that such content is regulated by default, regardless of users' personal configurations:

> *"I think that there are some things like public health misinformation that should just be automatically turned on. Like during COVID, there were a lot of hate crimes towards Asian people. I think platforms, especially large platforms, they know what's going on in the news cycle. They can curate certain moderation categories that are just automatically turned on because they know that misinformation's more likely to be higher in these spaces."* - P2

While participants wanted platforms to implement basic site-wide moderation that maintains community integrity and safety, they also wanted access to personal moderation tools to curate their individual experiences. For example, P6 and P9 noted that personal moderation tools would empower end users by giving them more control over what they see. P17 pointed out that the more specificity personal moderation tools allow users to dictate, the more helpful they would be. P22 considered it in platforms' interest to offer personal moderation tools because such tools let users shape their curation preferences while ensuring that platforms do not suffer any accusations of censorship.

We observed that many participants were willing to spend time and effort configuring their personal preferences for moderation. This readiness is likely because participants recognized that their personal preferences for content curation were niche and could not be captured by site-wide content curation systems. Some participants noted that platforms' site-wide policies and practices were often created in the context of Western culture and might not address the content curation needs of individuals in other cultures. This finding is in line with prior research showing that other moderation policies and mechanisms also often fail to account for localized contexts [50, 84, 103, 115]. Participants, therefore, considered it vital that personal moderation tools be offered so users can shape their own individualized, culture-aware preferences.

> *"I think it's a combination of both platforms and users [that should decide content curation] because otherwise the thing on the platform typically again is that it's going to be a team of majority white guys in some corner of the world deciding how an Indian boy is going to get his content."* - P10

While control over content curation is highly valued, most participants admitted that there should be restrictions on how personal moderation can shape their news feed. For example, there was a consensus that content that is politically neutral and beneficial for society, such as news of missing children or health information, should always be allowed, regardless of participants' personal moderation preferences.

Interestingly, many participants reported a lack of trust in platforms' motives, a sentiment primarily shaped by reading news reports about how company leaders have handled misinformation and online abuse. This loss of trust pushed some participants to seek out and rely on third-party moderation tools, especially those developed by universities or non-profit companies that are more positively perceived.

Many participants considered that in addition to platforms, end users, and third-party developers, lawmakers also have a role in ensuring appropriate news feeds. P15 thought governments must understand platforms' algorithms to process users' content. P14 and P21 feared that government intervention would increase bureaucracy and politically motivated censorship; they both proposed that a body of lawmakers from different countries be instituted to govern platforms. P21 said:

> *"Platforms need to be governed in the same way you would govern a sovereign nation, because they're almost an extension of our society, they're no longer just like a service that we use. I don't really trust the corporate structure to be able to police that system, just because they don't have that sort of incentive."* - P21

5 DISCUSSION

5.1 Addressing Harms while Still Enabling Freedom of Speech

5.1.1 Effectively Mitigating Potential Harms. Our analysis shows that individuals have different preferences regarding encountering offensive speech on social media. This observation is in line with prior research, which shows that while viewing extremist content on social media can be psychologically distressing for most, it is not experienced as harmful by everyone. Some even consider it an awareness-inducing experience that makes them feel better [97, 99]. Since users have varying moderation needs in response to the same content, platforms must prioritize the development of more advanced personal moderation solutions to facilitate individualized experiences and educate users about their utility.

Our participants' desire for such tools supports this call to action. As our participant responses show, currently offered moderation tools can be vastly improved by introducing search and analytics features that support easier configuration and selection. Prior research on moderation tools for content creators [52, 73] and moderators [13, 50] offers blueprints for designing and improving personal content moderation tools. For example, in line with the templates offered by FilterBuddy [52], category-based filters of the kind featured by Intel could be improved by configuring additional, carefully curated, culture-aware categories in partnership with trusted individuals or experts and allowing users to search for and select them. Thus, personal content moderation tools could address the challenge of incorporating cultural context in moderation, as highlighted by prior research [84, 103, 115]. We found that online harms are not limited to news feeds but can also occur in direct messages and user profiles (Sec. 4.1). Therefore, platforms must consider designing personal moderation tools that let users control their experience beyond news feeds. In fact, our analysis suggests that some users may prefer to use these tools to tune their search results but not their news feeds.

As our findings show, the fear of missing out (FOMO) makes users wary of over-moderating in response to harm. Thus, the challenge is to develop context-aware moderation solutions that understand users' values and safety requirements and then balance the competing needs. One analysis suggests that besides content removals, another promising approach is deploying content labels, such as fact-checks [118], interstitial warnings [14], or content sensitivity alerts, which can serve as reasonable middle-ground solutions to address users' complex needs [78].

5.1.2 Impact on Freedom of Speech. Personal moderation tools allow end users to control only what they see and do not restrict others' freedom of speech. In a few cases, participants muddied this distinction. Indeed, the broader use of these tools could *increase* free speech on platforms, as platforms may be more willing to leave borderline content up if users had an easy way to set personal preferences. Therefore, adding our voice to previous calls for educating users about moderation policies and related laws [49, 98], we recommend that users be informed about the possibility and bounds of personal moderation tools.

In some cases where participants spoke about "free speech," they objected to *other* users having the option to reduce their view of that content, i.e., they think everybody should be more open-minded. This discursive production of free speech as an argument against personal moderation tools expresses a moral position. It dictates the value of not shutting down certain viewpoints, but also, more crucially, requiring everyone else to do the same. This perspective suggests that even introducing optional personal moderation tools might engender some resistance.

A separate moral issue that specifically comes up in the case of slider controls is people's concerns with personal moderation controls that lets one see *more* of something offensive or harmful. Indeed, much of the outrage expressed online over Intel's Bleep tool focused on the fact that one could say they wanted the maximum level of some objectionable content. However, we note that this

is relative, with default placement potentially having a major impact on perception. In addition, framing what the slider accomplishes and the resulting slider labels for levels may also shift user opinion. Such framing may have been why Intel's sliders got significant pushback while a similar category-based 5-level slider feature on Twitch[6] for moderating chat messages in a Twitch stream did not. In the Intel case, one could specify, for instance, that they wanted "most" or "all" instances of aggression. In contrast, in the Twitch tool, the slider is flipped so that one can specify that they wanted "more" or "maximum" *filtering out* of aggression.

5.2 Informing Mental Models of Personal Content Moderation and Improving Controls

5.2.1 Clearer Definitions of Harm Categories. Our analyses show a crucial need for platforms offering personal content moderation tools to clearly and effectively communicate the meanings of their interface elements so that users can form accurate mental models of the type of moderation afforded and the degree of control they have. As we heard from participants, the use of generic, broadly defined keywords, such as 'toxicity,' 'moderation,' and 'name-calling,' raises questions in users' minds about whether their definitions aligned with those of the platform. Based on participant input, we suggest that platforms address this problem by offering more detailed transparency [94] into how they define and operationalize the moderation around abstract terms. For instance, if a platform decides to provide controls on moderating *toxicity*, they should offer additional information, like their working definition of toxicity and examples of what they consider toxic. In the same vein, when designing sliders that categorize content by type (e.g., "sexism", "racism") or along a scale (e.g., "very toxic," "somewhat toxic"), detailed information about what each category includes and how each level differs from others would be valuable. Further, when social or cultural trends necessitate updates in the system's operationalization of general moderation-related terms like toxicity, they should be communicated to the end users.

Of course, making definitions more grounded can be a fraught process since platforms would open themselves up to criticism from various individuals and communities who may have different expectations for what should be moderated. As Kumar et al. show, individuals frequently disagree on whether a comment is toxic and which subcategory (e.g., a threat versus an insult) a comment belongs to [64]. Platforms usually hide the massive machinery of content moderation to maintain a veneer of neutrality [83]. However, this secrecy ultimately works against them as it leads to suspicions that platforms are biased [114]. We argue that it is better to provide an imperfect but transparent moderation tool and hash out policy debates in the public than to provide a system that is opaque and unusable only to avoid controversy.

5.2.2 Transparency around the Algorithms Behind Controls. We found that in addition to definitions, users also desire more information about the inner logic and inputs to algorithms behind personal moderation tools. However, platforms usually keep this information opaque by claiming a need to protect their intellectual property [33, 98]. This information is necessary for users to avoid violating their expectations, which may contribute to a loss of trust [59]. Prior research has argued that instead of seamless interfaces, designing certain seams into an algorithmic system can affect users' understanding of that system and their interactions with it [12, 27, 54]. Thus, designers can experiment with building carefully designed seams, such as the main determining factors for moderation decisions, that expose more about algorithms behind personal moderation tools and examine users' responses.

While our findings show that offering visibility into moderation processes is crucial, this may require an overhaul of how personal moderation tools currently work behind the scenes. The internal logic of algorithms driving these tools is unknown, but their primary goal likely is to

[6]https://help.twitch.tv/s/article/how-to-use-automod?language=en_US

maximize toxicity classification accuracy, and they do not prioritize explainability. In the transition to making systems more interpretable, personal moderation tools would inevitably need to be more accountable. Prior research on related moderation mechanisms such as post-removals, flagging systems, and appeal procedures against moderation decisions have also called for platforms to be more accountable [17, 63, 106, 107, 114]. We add to these calls and offer specific recommendations for personal moderation tools. We suggest that designers alter the underlying algorithms at work so that their outputs are more understandable to the end user and in line with the definitional information provided by them to the greatest possible extent.

We found that participants often hypothesized about how personal moderation systems operated, forming their own (often inaccurate) folk theories [19, 54] about our simulated systems. Our analysis shows that inaccurate folk theories can lead to confusion and potential disappointment with personal moderation tools. While the full details of their inner logic may be too complex to be helpful to lay users, these tools do not even offer any *justification* [102, 110] for their moderation decisions by surfacing the key determining factors. Although the field of explainable models and AI is still emerging [2, 6], more can be done to provide details about the conventional aspects of systems. For example, platforms could clarify how moderation tools' overlapping settings interact, disclose whether systems are primarily rule-based or model-based, and reveal the criteria used during the training and evaluation of any models involved. Based on our participants' input, we suggest that using example comments at different stages is valuable to instill greater understanding. Providing additional transparency can help users decide which aspects of the systems they can trust and rely on and which parts need more caution.

5.2.3 Providing Desired Granularity and Continuous Adjustment. Furthermore, we found that while participants desired more control over personal moderation, platforms often failed to provide adequate controls to achieve that. For one, providing more degrees of freedom (like more levels of moderation) does not guarantee improved control over the system. Controls must be backed by a sufficiently crisp mental model of what the moderation system is meant to do and how well it can achieve that [47, 48]. Indeed, we found that the level of clarity in participants' mental models greatly affected how they engaged with personal moderation controls, with some noting that the number of slider levels should be *reduced* because they did not perceive a difference between some levels. Additionally, we noticed a strong desire for moderation tools to account for more context [4], like differences between communities and different linguistic patterns of groups and individuals. Designing controls to enable context-dependent moderation is a promising direction for future work. For example, interactive moderation tools that let users specify the relevant context (e.g., never hide racist comments from a specific user or group) would increase the utility of such tools.

Finally, we propose that configuring personal moderation be continuous and iterative. For one, the currently limited controllability of static personal moderation tools does not let users fully specify their moderation preferences. Additionally, some participants indicated that they may not be aware of their moderation needs until they encounter undesirable content. Prior research [52] has shown that users' moderation goals may also evolve over time. Therefore, platforms should offer mechanisms to adjust moderation in the context of results dynamically [48]. For instance, platforms should look towards providing just-in-time [37] controls, e.g., automatically incorporating the signal that "reported" content is undesirable so that users can customize personal moderation when it is most relevant to them. Our participants' enthusiasm for using examples as a form of control attests to the utility of this approach. Platforms should also provide feedback mechanisms, so users can systematically evaluate the efficacy of their current settings as opposed to the current ad-hoc reflections on their feed mentioned by our participants.

5.3 Managing Labor and Distribution of Responsibility

Our findings suggest that users keenly attend to the extent of labor required for personal moderation. Excessive cognitive demands of moderation configurations can deter users from engaging with personal moderation tools. Therefore, designers must devise *efficient* solutions that let users quickly configure their moderation preferences while retaining granular control. One recent example of such a solution is FilterBuddy, a word filter tool for YouTube that lets users quickly import entire pre-built categories of offensive keywords, e.g., about sexism and racism, but allow subsequent configuration changes for each phrase [52]. Given the tension between users' needs to minimize labor and improve control, designing tools that let users configure their preferences at varying levels of specificity offers a promising direction.

As our findings show, relying on like-minded others for co-creating shared moderation preferences can also reduce moderation labor. Further, showing improvement metrics in news feeds, e.g., the number of offensive comments hidden, can also encourage users to engage with moderation tools. Once appropriately configured, many personal moderation tools may need only occasional tweaks. Thus, platforms can emphasize the utility of a one-time investment in configuring moderation tools and make it easier for newcomers to understand and engage with them.

Regardless of how much platforms promote engagement with personal moderation, our analysis predicts that some users would hesitate to invest any time in changing their moderation settings, depicting a tendency to use the default settings [20, 93], while others may be unable to do so because of their lack of digital literacy. In addition to improving personal moderation capabilities, it is therefore vital that platforms provide sensible defaults. This recommendation is buttressed by our finding that most users expect a baseline level of platform review that would catch and remove blatantly inappropriate posts. However, platforms often fail to meet this expectation: our participants' frequent experiences of online harms speak to the urgency and importance of this effort. Thus, personal moderation tools do *not* absolve platforms of the necessity to conduct moderation. We argue against proposals that put all the onus on end users to personally filter content for themselves, as this may exacerbate the digital divide between those who can and cannot configure personal moderation settings.

In addition to platforms, third-party developers can also improve online spaces by implementing innovative moderation solutions. Our findings suggest that users are likelier to use tools built by academics and reliable non-profits. Moreover, policymakers can incentivize platforms to invest in creating new personal moderation interfaces or supporting the community of third-party developers. For this to occur, lawmakers require a better appreciation of what is at stake and a better understanding of how platforms implement personal moderation tools.

5.4 Limitations and Future Work

We focus on removing "toxicity" in three of our controls in our simulated social media feed. We chose this due to the availability of social media datasets with human-annotated labels for toxicity and the prevalence of the "toxicity" concept in academia and industry. However, personal content moderation can also include other types of content that may be undesirable to some, such as spam, sexually suggestive or explicit content, or violent content. It would be interesting to explore differing user preferences regarding these other categories of content and to understand the kinds of categories that users would like to adjust separately.

We deployed a toy web application that shows only 30 comments at a time, and it does not incorporate the effects of personalized content streams that users generally have within their social media feeds. Therefore, longitudinally studying user interactions with tools on real platforms hosting many more comments should reveal additional insights. The simulated controls we built

analyze only text, and our feed consists only of text comments. These controls could moderate other content, including audio, image, and video. The Bleep tool, as an example, focuses on filtering out audio in audio-based gaming chat rooms. Future studies could specifically examine personal moderation of audio, image, and video content, which may have different user preferences and considerations. For instance, violence and gore in visual content might be a higher priority for filtering since they may be more viscerally harmful. Our design exploration and toy website creation did not focus on a specific social media site but sought inspiration from various platforms. Since these tools are available alongside the platform-offered moderation mechanisms on each site, platform-specific explorations are needed to examine further the situated use and perspectives on personal moderation tools.

The specific terms (e.g., toxicity) that our toy platform uses could have influenced our findings about ambiguous definitions (Sec. 4.2.1). However, we note that Twitch and Instagram also use similarly abstract terms on their moderation sliders, such as 'offensive' and 'aggression,' that are likely to engender ambiguity. Future work that tests users on specific platforms would clarify the extent to which this problem persists in currently deployed tools.

Our method consisted of conducting semi-structured with 24 participants. While this let us elicit a wide range of concerns, user preferences, and nuances in engagement with personal moderation tools, the small sample size limited us from asking questions about the relative popularity of the different interfaces we tested. Going forward, we plan to conduct large-scale surveys with samples representative of the general population of internet users to answer such questions.

6 CONCLUSION

As a one-size-fits-all model for content moderation is insufficient, platforms need to consider the tools they provide end-users, so they can customize their moderation beyond what is caught at the platform level. We call this *personal moderation* and identify its two variations: *personal account moderation* and *personal content moderation*. We examine users' preferences regarding personal content moderation tools in this research. Our analysis shows that these tools would benefit from providing greater context awareness, clarity in the meanings of their interface elements, and justifications behind their decisions. Offering these tools does not exempt platforms from ensuring the efficacy of their baseline moderation. However, these tools can let users customize their social media experience without infringing on free speech concerns. Policymakers should also compel platforms to invest in building and supporting innovative personal content moderation tools.

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and color spaces.

REFERENCES

[1] Alfarez Abdul-Rahman and Stephen Hailes. 2000. Supporting trust in virtual communities. In *Proceedings of the 33rd annual Hawaii international conference on system sciences*. IEEE, 9–pp.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[3] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

[4] Gediminas Adomavicius and Alexander Tuzhilin. 2011. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 217–253.

[5] Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2016. Moodplay: Interactive mood-based music discovery and recommendation. In *Proceedings of the 2016 conference on user modeling adaptation and personalization*. 275–279.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence

(XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[7] Raéf Bahrini and Alaa A Qaffas. 2019. Impact of information and communication technology on economic growth: Evidence from developing countries. *Economies* 7, 1 (2019), 21.

[8] Jack M Balkin. 2013. Old-school/new-school speech regulation. *Harv. L. Rev.* 127 (2013), 2296.

[9] Bastien. 2021. Customized content moderation: One size doesn't fit all. https://www.bodyguard.ai/blog/customized-content-moderation-one-size-doesnt-fit-all

[10] Rahul Bhargava, Anna Chung, Neil S Gaikwad, Alexis Hope, Dennis Jen, Jasmin Rubinovitz, Belén Saldías-Fuentes, and Ethan Zuckerman. 2019. Gobo: A system for exploring user control of invisible algorithms in social media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 151–155.

[11] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. 2019. When users control the algorithms: values expressed in practices on twitter. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–20.

[12] Matthew Chalmers and Ian MacColl. 2003. Seamful and seamless design in ubiquitous computing. In *Workshop at the crossroads: The interaction of HCI and systems issues in UbiComp*, Vol. 8.

[13] 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.

[14] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *ACM Trans. Comput.-Hum. Interact.* (2022). https://doi.org/10.1145/3490499

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

[16] Samantha Cole. 2018. Where Did the Concept of 'Shadow Banning' Come From? https://www.vice.com/en/article/a3q744/where-did-shadow-banning-come-from-trump-republicans-shadowbanned

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

[18] Laura DeNardis and Andrea M Hackl. 2015. Internet governance by social media platforms. *Telecommunications Policy* 39, 9 (2015), 761–770.

[19] Michael A DeVito, Jeremy Birnholtz, Jeffery T Hancock, Megan French, and Sunny Liu. 2018. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 120.

[20] Nikhil Dhingra, Zach Gorn, Andrew Kener, and Jason Dana. 2012. The default pull: An experimental demonstration of subtle default effects on preferences. *Judgment and Decision Making* 7, 1 (2012), 69.

[21] Ana Diaz. 2021. Intel responds to hate speech tool getting roasted by the internet. https://www.polygon.com/22374120/intel-bleep-voice-chat-hate-speech-censor-spirit-ai

[22] Simon Dooms, Toon De Pessemier, and Luc Martens. 2014. Improving IMDb movie recommendations with interactive settings and filters. In *8th ACM Conference on Recommender Systems (Poster-RecSys 2014)*, Vol. 1247.

[23] Jack Dorsey. 2022. a native internet protocol for social media. https://www.getrevue.co/profile/jackjack/issues/a-native-internet-protocol-for-social-media-1503112

[24] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300372

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

[26] Maggie Engler. 2022. Middleware and the customization of content moderation. https://integrityinstitute.org/our-ideas/hear-from-our-fellows/middleware-and-the-customization

[27] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First i like it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. ACM, 2371–2382.

[28] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.

[29] Matthew Gault. 2021. Intel's Dystopian Anti-Harassment AI Lets Users Opt In for 'Some' Racism. https://www.vice.com/en/article/dyvgvk/intels-dystopian-anti-harassment-ai-lets-users-opt-in-for-some-racism

[30] R. Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803. https://doi.org/10.1080/1369118X.2016.1153700 arXiv:https://doi.org/10.1080/1369118X.2016.1153700

[31] Tarleton Gillespie. 2015. Platforms intervene. *Social Media+ Society* 1, 1 (2015), 2056305115580479.

[32] Tarleton Gillespie. 2017. Governance of and by platforms. *Sage handbook of social media. London: Sage* (2017).

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

[34] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media+ Society* 8, 3 (2022), 20563051221117552.

[35] Mike Godwin. 2003. *Cyber rights: Defending free speech in the digital age.* MIT press.

[36] Eric Goldman. 2018. The complicated story of FOSTA and section 230. *First Amend. L. Rev.* 17 (2018), 279.

[37] Damodar Y Golhar and Carol Lee Stamm. 1991. The just-in-time philosophy: a literature review. *The International Journal of Production Research* 29, 4 (1991), 657–676.

[38] Jay Graeber. 2023. Composable Moderation. https://blueskyweb.xyz/blog/4-13-2023-moderation

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

[40] F Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. 2015. Putting users in control of their recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 3–10.

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

[42] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56 (2016), 9–27.

[43] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[44] Silas Hsu, Kristen Vaccaro, Yin Yue, Aimee Rickman, and Karrie Karahalios. 2020. *Awareness, Navigation, and Use of Feed Control Settings Online.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376583

[45] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.

[46] Instagram. 2021. Introducing Sensitive Content Control. https://about.instagram.com/blog/announcements/introducing-sensitive-content-control

[47] Dietmar Jannach, Michael Jugovac, and Ingrid Nunes. 2019. Explanations and user control in recommender systems. In *Proceedings of the 23rd International Workshop on Personalization and Recommendation on the Web and Beyond*. 31–31.

[48] Dietmar Jannach, Sidra Naveed, and Michael Jugovac. 2016. User control in recommender systems: Overview and interaction challenges. In *International Conference on Electronic Commerce and Web Technologies*. Springer, 21–33.

[49] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 192 (Nov. 2019), 33 pages. https://doi.org/10.1145/3359294

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

[51] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action. *First Monday* 23, 2 (2018). http://firstmonday.org/ojs/index.php/fm/article/view/8232

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

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

[54] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems* (2018).

[55] Shagun Jhaver and Amy 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. (2023). In Preparation.

[56] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2022. A Trade-off-centered Framework of Content Moderation. *arXiv preprint arXiv:2206.03450* (2022).

[57] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PloS one* 16, 8 (2021), e0256762.

[58] Yucheng Jin, Bruno Cardoso, and Katrien Verbert. 2017. How do different levels of user control affect cognitive load and acceptance of recommendations?. In *Jin, Y., Cardoso, B. and Verbert, K., 2017, August. How do different levels of user control affect cognitive load and acceptance of recommendations?. In Proceedings of the 4th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2017)*, Vol. 1884. CEUR Workshop Proceedings, 35–42.

[59] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2390–2395.

[60] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.* 131 (2017), 1598.

[61] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*. 43–50.

[62] Joseph A Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User modeling and user-adapted interaction* 22, 1 (2012), 101–123.

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

[64] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. *arXiv preprint arXiv:2106.04511* (2021).

[65] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[66] Kalev Leetaru. 2019. Could personalized content moderation be the future of healthy social media? https://www.forbes.com/sites/kalevleetaru/2019/07/28/could-personalized-content-moderation-be-the-future-of-healthy-social-media/

[67] Lawrence Lessig. 2009. *Code: And other laws of cyberspace*. ReadHowYouWant. com.

[68] Clayton Lewis. 1982. *Using the" thinking-aloud" method in cognitive interface design*. IBM TJ Watson Research Center Yorktown Heights, NY.

[69] Jessica Litman. 1999. Electronic commerce and free speech. *Ethics and Information Technology* 1, 3 (1999), 213–225.

[70] Rebecca MacKinnon, Elonnai Hickok, Allon Bar, and Hae-in Lim. 2015. *Fostering freedom online: The role of internet intermediaries*. UNESCO Publishing.

[71] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friend-sourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[72] Tariq Mahmood and Francesco Ricci. 2007. Learning and adaptivity in interactive recommender systems. In *Proceedings of the ninth international conference on Electronic commerce*. 75–84.

[73] Keri Mallari, Spencer Williams, and Gary Hsieh. 2021. Understanding Analytics Needs of Video Game Streamers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[74] Mike Masnick. 2019. Protocols, Not Platforms: A Technological Approach to Free Speech. https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech

[75] Nathan J. Matias. 2016. The Civic Labor of Online Moderators. In *Internet Politics and Policy conference* (Oxford, United Kingdom). Oxford, United Kingdom.

[76] Aiden McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Cranefield. 2016. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. *ICIS 2016 Proceedings* (dec 2016). http://aisel.aisnet.org/icis2016/SocialMedia/Presentations/23

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

[78] Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P Wihbey. 2021. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology* (2021).

[79] Arvind Narayanan. 2022. How to train your TikTok. https://knightcolumbia.org/blog/how-to-train-your-tiktok

[80] Jonathan A Obar and Steven S Wildman. 2015. Social media definition and the governance challenge-an introduction to the special issue. *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy* 39, 9 (2015), 745–750.

[81] Denis Parra and Peter Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies* 78 (2015), 43–67.

[82] Block Party. 2022. Block Party. https://www.blockpartyapp.com

[83] Frank Pasquale. 2015. *The black box society: The secret algorithms that control money and information.* Harvard University Press.

[84] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '16)*. ACM, New York, NY, USA, 369–374. https://doi.org/10.1145/2957276.2957297

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

[86] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 317–355.

[87] Yolanda Linda Reid Chassiakos, Jenny Radesky, Dimitri Christakis, Megan A Moreno, Corinn Cross, David Hill, Nusheen Ameenuddin, Jeffrey Hutchinson, Alanna Levine, Rhea Boyd, et al. 2016. Children and adolescents and digital media. *Pediatrics* 138, 5 (2016).

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

[89] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation Accuracy Is Good, but High Controllability May Be Better. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

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

[91] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. *New Media & Society* (2020). https://doi.org/10.1177/1461444820913122

[92] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (2019), 1461444818821316.

[93] Rajiv C Shah and Christian Sandvig. 2008. Software defaults as de facto regulation the case of the wireless Internet. *Information, Community & Society* 11, 1 (2008), 25–46.

[94] Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*. 830–831.

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

[96] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, et al. 2022. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *arXiv preprint arXiv:2207.10192* (2022).

[97] Joshua Stubbs, Laura Nicklin, Luke Wilsdon, and Joanne Lloyd. 2022. Investigating the experience of viewing extreme real-world violence online: naturalistic evidence from an online discussion forum. (2022).

[98] 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.

[99] Sue Tait. 2008. Pornographies of violence? Internet spectatorship on body horror. *Critical Studies in Media Communication* 25, 1 (2008), 91–111.

[100] TL Taylor. 2018. Regulating the networked broadcasting frontier. In *Watch me play: Twitch and the rise of game live streaming*. Princeton University Press, Chapter 5.

[101] International Telecommunications Union. 2021. https://www.itu.int/itu-d/reports/statistics/facts-figures-2021/

[102] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*. Springer, 479–510.

[103] Rebecca Tushnet. 2019. Content moderation in an age of extremes. *Case W. Res. JL Tech. & Internet* 10 (2019), 1.

[104] Jean M Twenge, Gabrielle N Martin, and Brian H Spitzberg. 2019. Trends in US Adolescents' media use, 1976–2016: The rise of digital media, the decline of TV, and the (near) demise of print. *Psychology of Popular Media Culture* 8, 4 (2019), 329.

[105] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[106] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 167 (oct 2020), 22 pages. https://doi.org/10.1145/3415238

[107] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. 5, CSCW2, Article 318 (oct 2021), 28 pages. https://doi.org/10.1145/3476059

[108] José Van Dijck. 2013. *The culture of connectivity: A critical history of social media*. Oxford University Press.

[109] Maarten W Van Someren, Yvonne F Barnard, and Jacobijn AC Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress* 11 (1994).

[110] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*. 47–56.

[111] Ben Wagner. 2013. Governing internet expression: How public and private regulation shape expression governance. *Journal of Information Technology & Politics* 10, 4 (2013), 389–403.

[112] Galen Weld, Amy X Zhang, and Tim Althoff. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1121–1132.

[113] Barry Wellman, Anabel Quan-Haase, Jeffrey Boase, Wenhong Chen, Keith Hampton, Isabel Díaz, and Kakuko Miyata. 2003. The social affordances of the Internet for networked individualism. *Journal of computer-mediated communication* 8, 3 (2003), JCMC834.

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

[115] Richard Ashby Wilson and Molly K Land. 2020. Hate speech on social media: Content moderation in context. *Conn. L. Rev.* 52 (2020), 1029.

[116] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 160.

[117] Jillian C York. 2022. *Silicon values: The future of free speech under surveillance capitalism*. Verso Books.

[118] Savvas Zannettou. 2021. " I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter.. In *ICWSM*. 865–876.

[119] Mark Zuckerberg. 2021. Building Global Community. https://www.facebook.com/notes/3707971095882612/