Designing Word Filter Tools for Creator-led Comment Moderation

Shagun Jhaver
Rutgers University
New Brunswick, NJ, USA
shagun.jhaver@rutgers.edu

Quanze Chen
University of Washington
Seattle, WA, USA
cqz@cs.washington.edu

Detlef Knauss
University of Washington
Seattle, WA, USA
detlefk@uw.edu

Amy Zhang
University of Washington
Seattle, WA, USA
axz@cs.uw.edu

ABSTRACT

Online social platforms centered around content creators often allow comments on content, where creators can then moderate the comments they receive. As creators can face overwhelming numbers of comments, with some of them harassing or hateful, platforms typically provide tools such as word filters for creators to automate aspects of moderation. From needfinding interviews with 19 creators about how they use existing tools, we found that they struggled with writing good filters as well as organizing and revising their filters, due to the difficulty of determining what the filters actually catch. To address these issues, we present FilterBuddy, a system that supports creators in authoring new filters or building from pre-made ones, as well as organizing their filters and visualizing what comments are captured by them over time. We conducted an early-stage evaluation of FilterBuddy with YouTube creators, finding that participants see FilterBuddy not just as a moderation tool, but also a means to organize their comments to better understand their audiences.

CCS CONCEPTS

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

KEYWORDS

platform governance, YouTube, online harassment, content moderation, content creators, human-computer integration, FilterBuddy

ACM Reference Format:
Shagun Jhaver, Quanze Chen, Detlef Knauss, and Amy Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3491102.3517505

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517505

CONTENT WARNING: This paper contains offensive language, including misogynistic slurs, that readers may find disturbing.

1 INTRODUCTION

> *"Some of you have never had to schedule dedicated time each month in your calendar to remove all the 'nice boobs'/'she's cute'/'can she actually code?'/'get back in the kitchen' comments from your technical videos on YouTube, and honestly? It shows. [Please] do not comment telling me to turn off comments…I like to get comments/nice things said to me (just like anyone else would)…And this is just YouTube 🙃 Don't forget Vimeo/my Instagram/Twitter/LinkedIn/everything else, too!* 💀🪶🪦👻
> – Chloe Condon, 2021 [20]

Platforms like YouTube, TikTok, and Twitch combine technical and social features to support the formation of a participatory culture in which ordinary individuals create, collaborate, and share personally meaningful media, often as a source of income [18, 81, 93]. With thousands of new creators joining these platforms every year, the competition among creators for attracting viewers is high [47]. In addition to the content they produce, many creators distinguish themselves by working to curate a vibrant and inviting community around their content through interacting with and moderating their audience's commentary.

However, as Chloe Condon, a content creator who makes technical tutorials on YouTube and elsewhere, addresses in her Twitter thread excerpted above, creators—in particular, ones from marginalized groups—must invest a great deal of emotional labor to address unwanted comments from their viewers [86]. As channels grow larger and attract more comments, it becomes increasingly difficult for creators to review every comment, and some automation becomes necessary. Even for the small minority of creators who can hire staff or are supported by volunteer moderators, receiving abuse via comments can contribute to anxiety and symptoms of post-traumatic stress disorder for creators due to the oftentimes personal nature of the attacks [86]. In light of this, creators like Condon have frequently requested more powerful moderation tools and resources that can meet the needs of marginalized creators and alleviate the pressures surrounding content moderation.

In this paper, we investigate the design of moderation tools aimed at content creators to understand how they can better address creators' needs, especially the needs of creators from marginalized groups who face disproportionate and targeted abuse [40]. We focus our needfinding and design exploration around *word filter tools*, one of the most common moderation tools that allow creators to configure a list of phrases such that comments containing any of those phrases get automatically removed or held for review. Through needfinding interviews with 19 creators, we examine their experiences with receiving unwanted comments and their everyday practices and strategies using existing word filter tools to try to reduce their moderation workload and stress. We found that creators were overall frustrated with the rudimentary features provided by existing moderation tools on platforms such as YouTube. Creators described having difficulties with both building up a set of useful filters from scratch, as well as organizing a growing list of filters and auditing what their filters were actually catching.

To address these needs, we present FilterBuddy,[1] a system built for YouTube creators that augments word filter tools with features to support better authoring, maintenance, and auditing of word filters. Users can connect the system to their YouTube channel and use it instead of the native word filter tool on YouTube to moderate their comments. Features of FilterBuddy include but are not limited to: interactive previews of what different filters would capture during the authoring process; the ability to build from existing filters created by others; organization of filters into spelling variants and higher-level categories; and time-series graphs and tables to understand which comments are caught by which filters over time. FilterBuddy is not just a research tool; it is a system designed to be actively used by YouTube creators to enhance their moderation capabilities. While this tool was designed to address the needs of marginalized creators and protect against online harassment, we have found that it may also have a wider range of uses and could benefit a broader range of creators.

We conducted an exploratory qualitative user study of FilterBuddy with 8 creators on YouTube who interacted with the tool loaded with their YouTube comments while providing feedback. We found that participants were appreciative of greater automation but did not prefer to sacrifice their control over the tool's operations and requested additional defensive mechanisms to reduce incorrect removals. They also considered the ability to share their filters and build off of filters created by others to be a powerful means to reduce toxic content across YouTube. In addition, we were surprised to find that participants devised many use cases for the tool that go beyond just moderating undesirable posts. Overall, participants felt that the designs explored in FilterBuddy would empower content creators, especially those belonging to marginalized groups, to efficiently address their content moderation needs as well as better understand their audiences.

We conclude by discussing the tensions and tradeoffs in creators' goals that would be important to consider when designing rule-based moderation tools. Our design exploration shows that creators are deeply invested in retaining control over their moderation operations; we reflect on how sensible design defaults can help creators achieve this goal while minimizing the manual effort of setting up granular filters. We also highlight that trust-sensitive collective governance mechanisms will be required to resolve the tensions creators face between seeking to build on other creators' word filters while also preferring to keep their configurations private to avoid exploitation by bad actors. Finally, we had limited development resources to implement FilterBuddy—yet, we found that its features are seen as highly desirable by creators. In light of this, we call upon platforms to step up their efforts and investments in developing tools that improve creators' working conditions. We also examine how other stakeholders such as policymakers, third-party developers, and minority support groups can offset the moderation workload of creators.

[1] https://filterbuddy.org

2 RELATED WORK

We focus our review on content creators and the unique challenges they face, the nuances and difficulties of addressing online harm and safety, and the design of existing end-user focused content moderation tools such as word filters.

2.1 Experiences of Content Creators Online

Rajendra-Nicolucci and Zuckerman define *creator logic platforms* as platforms that "enable users to share a specific type of media (like video, livestreams, or art), in a one-to-many fashion." [93]. Such platforms include YouTube, TikTok, and Twitch, which let people from around the world upload content online and accumulate relatively large followings [1, 43]. Dominant social networks like Facebook, Instagram, and Twitter are used for a broader variety of purposes, such as connecting with family and friends, but creators can use these networks to share their content with a large audience as well [93].

Content creators not only have to write, produce, and edit their content, but they must also cultivate and engage with their audiences [42, 45, 62, 125]. Creators often adopt strategies of "micro-celebrity," where they regard their audience as fans, perceive content delivery as a performative act [67, 90], and use strategic intimacy to appeal to viewers through direct interactions [3, 73, 94]. The technical affordances of content sharing platforms, such as the ability for users to view, comment on, and share content, also encourage the development of active, networked groups [1, 23, 64], where content creators and viewers interact with each other [6].

As a result, creators often find themselves overwhelmed with the varied demands of their work and can suffer occupational stress [86]. While a small minority of content creators become successful enough to hire content moderators, the vast majority do their best to manage different tasks by themselves [125]. Opaque algorithms used by these platforms evaluate content creators and shape their popularity and earnings, thereby rendering consistent earnings uncertain and risky [89]. As a result, creators often hesitate to hire staff for help even though platforms encourage struggling creators to enlist support [86]. Platform-enacted demonetization of creator posts that contain inappropriate comments are an additional concern for creators when conducting content moderation [5, 10]. Many creators also experience doxxing, harassment, stalking, and online threats, which further contribute to their mental health risks [66, 114]. Prior research shows that existing platform tools fail to address a range of creator needs, including community information

needs [72]. In light of this, any tool that helps creators manage their channel can benefit not just the creators by reducing their workload and stress but also improve the experience of their audiences.

2.2 Online Harm and Safety in HCI

Online harm and abuse are highly subjective, situated concepts [101] whose interpretations vary across cultures [59] and even across individuals [55]. Online platforms do not always explicitly define these concepts in their policies and oftentimes take an ad hoc approach to setting standards for addressing them [60, 88]. Scheuerman et al. identified four types of online harm—physical, emotional, relational, and financial [102]. Elsewhere, Scheuerman et al. showed that experiences of online harm can occur in varied ways, i.e., abuse aimed directly at the individual or witnessed by the individual; abuse perpetrated by those outside or within the individual's social circles; and abuse impacting a specific individual or larger communities [101]. Content creators and their viewers can experience harm in all of these ways through comments posted on their channel.

The related concept of *safety* is usually referred to as protection from emotional, physical, and social harm that may or may not be caused by abusive behavior [44, 61, 96, 101, 117]. Prior HCI research on safety has largely focused on interpersonal harm, which includes hate speech, cyberbullying, and online harassment [9, 69, 76, 92, 107], and how it causes emotional distress or threatens physical safety [101]. Redmiles et al. uncovered the multifaceted nature of online safety on Facebook, arguing that it involves not only digital privacy, digital security, and harassment, but also offline safety and a sense of community support as well as the upholding of community values [95]. Building upon this research, we examine the safety perceptions of content creators and explore how to design an anti-harassment tool that is privacy-sensitive and that contains features to foster community support in order to increase the safety of content creators online.

One particularly distressing statistic related to online harm and safety is that marginalized communities are disproportionately affected by internet-facilitated harassment and cyber-bullying [57, 71]. For example, in 2020, Blacks (54%) and Hispanics (47%) were more likely than Whites (17%) to report victimization due to their race/ethnicity, and 47% of women claimed to have been harassed online due to their gender as opposed to 18% of men [118]. Prior research has also documented incidents of online abuse against members of the Black Lives Matter movement [51], women [29, 83], and LGBTQ communities [32, 101]. Such disproportionate abuse is also reflected in the experiences of content creators, e.g., female and LGBTQ Twitch streamers frequently encounter harassment [34, 114]. Our needfinding interviews with creators from marginalized groups echoed these findings, and we built on this prior work to explore how creator-led moderation tools can empower creators from marginalized communities to resist interpersonal harm.

HCI researchers have also focused on content-based harm, which occurs as a result of viewing undesirable content on social media platforms. For example, this includes race-based trauma caused by viewing posts of police violence against Black people [111, 112, 123]. Research on the work practices of content moderators has shown how continuous exposure to violent, hateful, or otherwise troubling posts takes an emotional toll, causes secondary trauma, panic attacks, and other mental health issues, and eventually burns out many workers [26, 53, 97, 108, 124]. We describe in this work the kinds of content that creators want to filter out and present a system aimed at reducing the need to continually expose oneself and one's audience to content-based harm.

2.3 Content Moderation Tools for End-Users

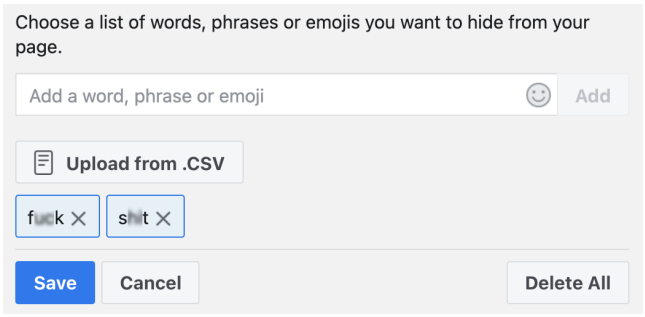
Content moderation refers to a series of practices, including rule-setting, using human moderators, and deploying automated tools, that are used to screen user-generated content to decide what posts will appear on, or remain on, a platform [11, 36, 97]. On many platforms, including Reddit, YouTube, and Twitch, content moderation is performed at multiple *levels* of governance [56], where content creators and volunteer community moderators have significant leeway to make and enforce local rules regarding comments [53, 58, 77, 124]. Meanwhile, the platform oversees automated takedown of certain classes of posts globally and also responds to user reports. In some cases, the platform may also step in to override an individual creator or community's self-policing [11, 14, 15, 48, 54].

When it comes to creator-oriented platforms in particular, much of the work of removing inappropriate comments on one's content falls on the shoulders of content creators. Platforms are also responsible for providing tools for creators to be able to perform moderation adequately; however, these tools can fall far short of addressing users' moderation needs. For example, Matias describes how the prolonged neglect of moderation software on Reddit was so frustrating that it pushed moderators of 2,278 communities to join together in protest against the platform [75]. Our work contributes to a growing thread of content moderation research that focuses on tool design and techniques targeted at end users to support them in countering harms like online harassment [4, 50, 57, 71, 117].

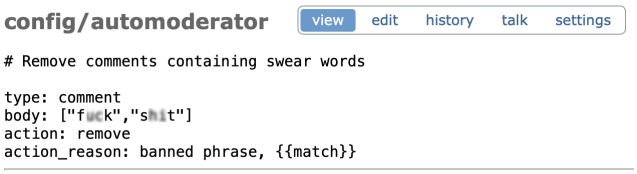
2.3.1 User blocking and reporting. When it comes to the techniques that platforms have implemented, one common end-user moderation technique is reporting a post as inappropriate to the platform [21, 63]. However, some critics have pointed out that the current systems' reliance on reports of dangerous or offensive content by users constitutes a reactive, too-little-too-late moderation approach that fails to protect the already-offended users [80, 106, 113]. As an example, a cross-platform study of queer female users showed that reporting to platforms failed to protect them from harassment and discrimination [27].

Another approach is blocking accounts from commenting or viewing one's profile [8, 58]. While effective at small scale, it can be circumvented by motivated harassers creating new accounts [71] and is less attractive as an option in the face of large and changing audiences, as is the case sometimes for content creators. Some third-party tools [46, 87] address these weaknesses by allowing users to configure rules and share lists to automatically block accounts. However, these approaches can sometimes inadvertently block would-be commenters that are well-meaning, a concern for creators focused on growing their audience. Thus, we draw inspiration from these tools while focusing on moderation that examines content.

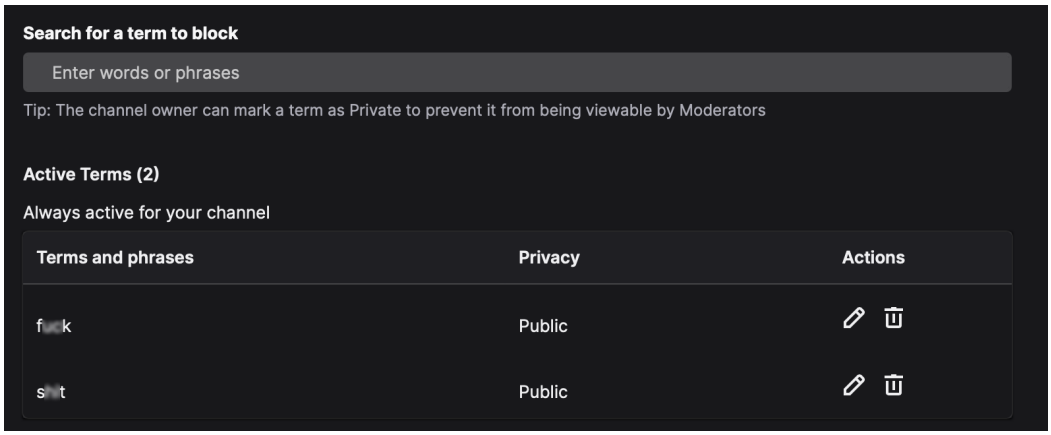
2.3.2 Word filter moderation tools. When it comes to content-based approaches, one of the most common types of moderation tools



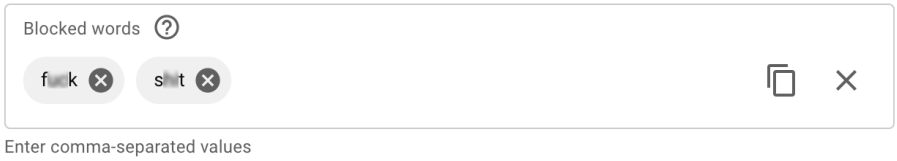
(a) Facebook Page



(b) Reddit Automoderator



(c) Twitch



(d) YouTube Studio

Figure 1: Implementations of word filters on four popular platforms—Facebook, Reddit, Twitch, and YouTube. (a) This setting appears on Facebook Pages, which are public Facebook profiles that allow "businesses, brands, celebrities, causes and organizations to reach their audience." [7], (b) this community-specific setting is accessible only by the volunteer moderators of each Reddit community [53], (c) this setting is available to Twitch creators on their *Creator Dashboard*, a webpage that allows creators to organize content and improve safety preferences [22], and (d) this is available as part of *Community Settings* on YouTube Studio, an official YouTube site where creators can grow their channel, interact with audiences, and manage earnings [109]. In each case, the shown interfaces are the only information sources available for viewing and configuring word filters; no accompanying feedback or visualization mechanisms exist to assist configuration or track how the filters are performing.

available to end-users, and the focus of this paper, is *word filters*, which lets users configure a list of blocked keywords. Once a keyword is configured in a word filter, any post containing that keyword is automatically removed or held for manual inspection. Our review of moderation tools across popular social network sites found that word filters are offered by YouTube, Twitter, Facebook Pages, Reddit, Instagram, and Twitch, among others. Word filters are a subset of rule-based moderation tools, or tools that allow end-users to author executable rules over comments to determine what to filter out. Rule-based methods are currently in use by many online communities [53, 58, 104]. For example, Reddit moderators can configure automated removal of inappropriate comments using rules that specify `<regular expression, action>` pairs (see Figure 1b) in a subreddit-specific tool called Automod [53].

As online communities or channels grow to a size that human moderation can no longer handle, rule-based moderation tools offer a hopeful solution for enacting moderation at scale. These tools can reduce the time-consuming work and emotional labor required of

human moderators by automatically removing large volumes of inappropriate content. However, we note that current word filter interfaces are relatively rudimentary—little more than text boxes for inputting comma-separated terms—with few features that incorporate interactivity, visualization, or affordances for information management (see Figure 1). Sometimes they are configured using regular expressions, which are difficult to set up and require moderators to develop new technical skills. Prior research has also highlighted the need for audit tools that provide moderators with greater visibility into the performance of such tools [53]. In this paper, we examine in detail the current use of word filter tools, identify the problems that creators face in using them, and contribute a system to address those problems.

2.3.3 Machine learned approaches. Finally, researchers have also examined machine learned approaches for identifying hate speech [16] and personal attacks, as well as high-quality comments [85]. Some platforms incorporate machine learned tools that are targeted at creators. For example, Twitch creators can configure an automated moderation tool called Automod to select one of five levels (including the lowest level that turns off Automod) that affects how aggressively AutoMod prevents messages under four categories offered by Twitch: discrimination and slurs, sexual content, hostility, and profanity [79].

While machine learning has frequently been proposed as a wholesale solution to proactively remove undesirable posts, it requires gathering enough training data on rule violations, may not be adaptive to new kinds of infractions [37, 58], may further complicate fairness and justice issues [38], and may have decisions that are not comprehensible [53]. Recent research suggests that algorithms designed by software engineers to detect hate speech show biases, e.g., they remove racial slurs but not gendered slurs [11, 82, 99]. In this work, we focus on improving systems for rule-based automation of content moderation, finding that creators value them for being controllable and understandable, and we discuss the trade-offs with more sophisticated automated techniques.

3 NEEDFINDING STUDY: EXPERIENCES, CONCERNS, AND STRATEGIES

We began with a needfinding study to understand the current moderation practices of content creators, including their strategies and challenges with using common tools like word filters.

3.1 Method: Interviews with Content Creators

3.1.1 Interviews. We conducted semi-structured interviews with 19 content creators on sites like YouTube, Twitch, and TikTok between October 2020 and April 2021. We first prepared a *semi-structured interview script* to understand the moderation practices and word filter usage of interview participants. As we gained more understanding of the fundamental moderation issues, we continued revising our interview script to elicit deeper insights. We began by asking questions about general social media use as a content creator to understand the context in which participants work. This included questions about the platforms used to upload content, how and why participants review comments on their content and engage with their audiences, and the types of inappropriate comments they receive. Next, we explored how and why participants began to use word filters, the changes they observed as a consequence, and the challenges they faced in configuring these filters. Further, we queried about the design features and changes they would like to see incorporated into these word filters.

Based on a few initial interviews, we began designing *multiple alternative solutions* for serving three user needs: (1) capturing the meaning of undesirable ideas, (2) understanding which posts are caught by each word filter configuration, and (3) specifying actions to take on word filters. For example, to capture the meaning of undesirable ideas when entering a keyword, we presented two design sketches: (a) auto-suggesting similar additional keywords that participants may want to filter and (b) auto-suggesting entire categories of offensive keywords such as antisemitism and transphobia. To take another example, when setting up a word filter, we presented two ideas for configuration assistance: (a) showing examples of recent comments matching that filter and (b) showing descriptive statistics (e.g., number of comments, number of users, number of comments by the channel's moderators) of comments caught by that filter.

We described these solutions to participants and asked them to reflect on the solutions. Asking participants to contrast alternative solutions as shown above helped us elicit more creative and detailed feedback. Most participants wanted a single tool offering multiple informational and visualizations features serving their different needs; we therefore attempted to design a tool that offers multiple features but does not cognitively overwhelm the user. We constructed *preliminary design sketches* using the tool Figma to incorporate these features; we sought participant feedback on these sketches and continued revising the sketches based on participant responses, particularly ensuring that the interface is easy-to-use. Our evolving design ideas and sketches led us to ask many additional questions addressing word filter categories that participants would like to have by default, and how participants felt about sharing their word filter configurations with other creators.

3.1.2 Participants. We adopted convenience sampling to recruit our subjects. We posted recruitment messages on social media sites and sent personal messages to participants we had interviewed in our previous studies on related topics. We sought to recruit a diverse set of participants, ensuring that our sample included creators with channels of different sizes who were focused on a variety of topics and reside in multiple countries. Prior research has shown that BIPOC, LGBTQ+, and female users have experienced higher instances of online harassment and harms of content moderation [41, 115]. We therefore oversampled gender, racial, and sexual minorities since we expect that these groups would especially benefit from upgraded moderation tools. Our selection criteria included choosing only those creators who had received at least a few dozens of comments on their channels so that they have more experience with the challenges of content moderation at scale. We also selected only those participants who had either used YouTube Studio's *Automated Filters* or other similar automated ways to manage comments on their videos since our interviews dealt with automated aspects of comment removals. Table 1 presents demographic information about our participants and the size and topical focus of their channels. To maintain privacy, we do not reveal participant names or channel details in this paper.

Table 1: Demographic information of needfinding study participants. Here, 'Platform Used' includes all the platforms where participants share the content they created. 'Size' refers to the highest number of subscriptions participants received on their channel on any platform; for participants with size 'N/A', the participant did not divulge the size of their channel. 'Topic' refers to the main topical focus of creator's channels.

Sr. no.	Age	Gender	Occupation	Country	Platforms Used	Size	Topic
P1	28	Male	Banker	USA	YouTube, Reddit	>100k	American football
P2	24	Female	Youtube Accounts Manager and Producer	Jordan	YouTube, TikTok, Facebook, Instagram	500-1k	Family life
P3	23	Female	Student	USA	YouTube	20k-50k	School life
P4	19	Female	Student	USA	YouTube, TikTok	<500	Music creation
P5	20	Male	Student	India	YouTube	<500	Gaming
P6	21	Male	Student	Germany	YouTube, TikTok	N/A	Gaming
P7	24	Male	Student	India	YouTube	<500	Technology news and advice
P8	20	Male	Student	India	YouTube, Reddit, Facebook, WhatsApp	N/A	Technology
P9	28	Male	Student	USA	YouTube, Facebook, Instagram	<500	Satire
P10	29	Nonbi-nary	Freelance Creator	USA	YouTube, Instagram	20k-50k	LGBTQ
P11	28	Trans Female	YouTuber	Australia	YouTube	>100k	Gender identity
P12	29	Female	Communty Manager	USA	YouTube, Twitch	5k-10k	Miscellaneous
P13	23	Male	Student	USA	Twitch	N/A	Gaming
P14	31	Female	Biologist	Brazil	YouTube, Instagram	10k-20k	Science
P15	35	Female	Content creator	USA	YouTube, Twitch	500-1k	Gaming
P16	27	Female	Content creator	UK	YouTube, Twitch, Instagram	20k-50k	Makeup
P17	35	Male	Podcaster	Brazil	YouTube, TikTok	10k-20k	History
P18	29	Female	Customer Service Representative	UK	Twitch, Instagram	500-1k	Art
P19	28	Male	Student	Brazil	YouTube, Instagram, Twitter	1k-5k	Science

The interviews lasted between 75 to 90 minutes and were conducted through video chats using Zoom or Skype, depending on each participant's preference. Most participants used YouTube as their primary platform to upload content, but some used other sites like Twitch and TikTok to share and stream videos. All interviewees received $25 as compensation for their participation.

3.1.3 Analysis. We fully transcribed our interview data and read them multiple times. One interview was conducted in Hindi; we translated it to English to simplify interview coding. Next, we applied interpretive qualitative analysis to all interview transcripts [78]. This involved a rigorous categorization of our data to identify relevant patterns and group them into appropriate themes. We began with "open coding" [17], for which we assigned short phrases as codes to our data. We conducted this first round of coding on a line-by-line basis to stay close to the data. Next, we engaged in multiple subsequent rounds of coding and memo-writing, conducting a continual comparison of codes and associated data. All authors discussed the codes and emerging concepts throughout the analysis and resolved conflicts through discussions. In later rounds, we began combining our initial codes into high-level ones, such as "Prefers to handle comments automatically" and "Wants to see statistics." Once all authors agreed upon the codes, we distilled them into key themes and performed axial coding to deduce relationships. We next present these themes as our findings.

3.1.4 Maintaining Participants' Privacy and Security. Content moderation is a sensitive topic and creators, owing to their online public presence, are especially vulnerable to retaliation by internet trolls. Therefore, we took many important steps to ensure the privacy and security of our participants. We anonymized all our participants' identities in this manuscript and plan to continue doing so for all future presentations and demonstrations of this work. All comments and usernames shown on screenshots of our tool in this paper are synthetic. We also stored our interview recording and transcripts (for both need-findings analysis and the later user evaluation) on a secure, password-protected server.

3.2 Findings: Moderation Concerns and Strategies

We now describe content creators' experiences with receiving offensive comments, the moderation strategies they use to improve their content feeds, and their concerns with attaining balance in content moderation.

3.2.1 Creators value engaging with user comments, but find frequent, undesirable comments disruptive. All study participants indicated that they value reading user comments. Five participants whose channels focus on social justice issues or spreading public awareness about science and history pointed out that reading comments helped them realize how they are contributing to their audience. For example, P2 manages a YouTube channel that hosts videos about socially relevant issues and noted:

> "*I actually enjoy reading comments because they give you a sense of achievement, and it's really rewarding honestly, especially on videos where it's building awareness and seeing how people are reacting to that. For example, we did a video about gender equality, and it was really, really nice to see how people are reacting to that and their own experiences.*" - P2

However, viewing and engaging with comments become difficult when they are offensive, spam, or off topic. Echoing prior work, we found that participants felt particularly disturbed when these comments were harassing [66, 114]. For example, participants from certain identity groups (including users who were Asians, Blacks, LGBTQ, Muslims, female, or of Middle Eastern descent) reported experiencing identity-based attacks that severely distressed them. Two participants also reported suffering from organized harassment or waves of negative comments that were emotionally distressing. P3 also expressed concern about her audiences viewing offensive comments:

> *"It feels almost kind of embarrassing for someone to say like, "You're ugly," or, "You're dumb," and then for other people to see that comment. I want to form a positive community, and if I don't control those sorts of language, then people will think it's okay, and then slowly the community will become a lot more negative overall." - P3*

P16 noted that as channels become more popular and attract more viewers, the probability of receiving offensive comments also increases. In one case, such harassment caused the participant to have anxiety attacks, seek medical help, and subsequently stop posting videos on his channel. Female, transgender, Muslim, and BIPOC users were particularly bothered by receiving comments on their appearance. For instance, P14 talked about her experience of having a transgender co-creator on her science-based channel:

> *"[We receive comments about] that friend of mine that is a trans woman about what is she doing, what is she wearing, why is she talking like that, why is she there? But she also is a scientist, she also has a Ph.D., and she is an amazing paleontologist." - P14*

3.2.2 Creators use a variety of strategies to moderate. While nine participants stated that they often just ignored offensive or off-topic comments, others felt the need to engage in a variety of moderation practices to curb harassment both for themselves and their viewers. In the most extreme cases, participants disabled the comment section entirely, restricted the visibility of their account or video, or even removed the video. For example, P10 stated:

> *"The only way to stop [the harassment] from happening was... to make my account private and not let other people who are trying to come at me follow me." - P10*

More commonly however, participants elected to use their platform's built-in word filter tool to hold comments for review or block commenters. In addition, most participants also manually moderated their comments to some extent. Five creators replied to offensive comments to try and educate commenters and encourage better behavior. Some participants, especially those who streamed on the Twitch platform, used the help of volunteer moderators to manually moderate their comment sections as well. Those participants usually had a set of guidelines for their moderators to follow:

> *"It's different per channel. What I think is acceptable or is unacceptable may be different to another creator so... you tell your moderators what is acceptable and what isn't." - P16*

3.2.3 Creators may receive backlash for their moderation practices. While creators struggle most often with sufficiently moderating their comments, five participants reported receiving backlash in response to moderation practices perceived as too strict. They told us that commenters reacted poorly if they felt that moderation was overly restrictive. P3 noted:

> *"I've seen posts on Reddit or even just on someone's Instagram page that's like, 'Oh she's removing comments,' or, 'This person is removing hate comments,' or, 'Why's my comment not appearing?' It's kind of like accusatory." - P3*

Moderation is seen as striking a balance between *too much* and *too little*. Under-moderation results in harassment and abuse in the creator's comments section. Over-moderation, though, can drive viewers away from the creator's channel, possibly costing the creator financially, or even cause commenters to harass the creator more, as Participant P11 pointed out:

> *"I think [the comments] would get caught in the word filter and then they wouldn't show up on the video. And so, people would think that I'm manually deleting them, which would then... encourage them to get more angry." - P11*

Summary. In summary, creators value engaging with user comments, but frequently encounter inappropriate comments that disrupt this process. To address this problem, creators try to educate offenders, restrict content visibility, and get assistance from volunteer moderators. They see content moderation as a balancing act between preventing abuse and retaining audiences.

3.3 Findings: Current Use of Word Filters and Need for Improvements

Next, we focus on the content creators' use of word filters, one of the most prominent moderation tools available to creators on many platforms. We examine the current practices of using these filters and surface crucial needs that are not currently met by this tool. Specifically, we explore what additional informational elements and visualization features content creators would like to have in this tool, how they would benefit from more automation, and their attitudes towards sharing word filters they configure with other creators.

3.3.1 Incorporating a greater degree of organization. 15 participants wanted to create different categories of word filters, such as racism, sexism, etc., to separately group different sets of inappropriate comments. For example, Participant P2 felt that having multiple categories would allow her to better organize and contrast comments caught by different categories as well as configure separate sharing preferences for each category. Participants P4 and P7 desired multiple categories because they wanted to configure different actions for each category. Five participants noted the utility of separating comments into distinct groups for reasons beyond moderation, such as easier review and search.

> *"Well, we do it manually—we read every single comment that we get. And sometimes we get thousands and thousands of comments... So it would take so much less time and effort to just have them categorized into...a category where you can just see what sort of comments that you're getting and the percentages of good to not so great." - P2*

3.3.2 Assistance with configuring keyword spelling variants. When asked how they use these word filters, most of our participants noted that they configure them with curse words. However, eight participants experienced commenters evading and bypassing word filters to get their offensive comments read. This could involve using variations in the spelling of a slur or offensive word, spelling out offensive words phonetically or in a different language, or many other methods. To address these problems, nine participants wanted the ability to more easily configure similar words or creative misspellings. At present, this would require the creator to either spend considerable time configuring all possible variations of a slur or to spend more time manually moderating and removing comments. When presented with a design where spelling variants are automatically generated when adding a slur to a word filter, P10 responded:

> *"It takes a lot of the burden of having to ideate all of this off of the creator...without this, that's a real emotional moment of labor of having to again, manually input all these things and create these really long block lists." - P10*

3.3.3 Understanding what different configurations would capture. One of the most common problems that creators related about their use of word filters was that the filters catch many false positives. Frequently, built-in word filters flag innocuous, even valuable, comments and bury them in review folders that are full of genuinely offensive or spam comments. P2 said:

> *"Sometimes, like in the last few months, I would post a video and then like an hour after, I would just go check the spam filter, and there would just be people talking about normal stuff...There's probably a bunch of nice comments caught in there but I don't really have the brain space to go back and look through them and approve all of them, so I just kind of leave them there, which sucks."* - P11

These issues resulted in three creators choosing to not use word filters at all. For example, due to the frequency of false positives, P15 pointed out:

> *"I have [the word filter] disabled right now because it was being very excessive." - P15*

In response to these issues and limitations, nine creators desired to see examples of matching comments when configuring keywords in word filters. They believed this would enhance both their understanding of how the word filter would perform and how to improve its performance. For example, P7 said:

> *"[Seeing examples of matching comments] is interesting for me, because you can understand—because in a sentence, you can have multiple meanings of it, because even the word 'alpha'... it's a smartphone name as well. So, if it is included in the sentence, you could mean a lot more because I can understand if they're talking about the phone or if they are slurring at me." - P7*

3.3.4 Importing built-in categories. Black, female, and Muslim participants expressed a need for word filters to have a few default curated categories, like racism and sexism, that they can opt into so that they do not have to create such commonly needed categories from scratch. P15 noted:

> *"If you guys created some sort of like—'we've gathered together a panel of these five people to create our lists for this year and then every year it was edited'—right, because culture is constantly evolving, then I'd know that you guys have taken the time to select each of these people, and then I'm more likely to trust you guys as thought leaders in the space and use the lists."* - P15

3.3.5 Getting an overview of word filters' actions. In addition to seeing what individual word filters would capture at the time of configuration (Sec. 3.3.3), creators were also interested in reviewing the overall temporal performance of configured categories. For instance, twelve creators wanted to see a summary of actions taken by word filters, e.g., descriptive statistics about a filter's performance, to catch issues like over-moderation. P9 noted:

> *"Obviously, something like adding word filters could potentially have a negative impact if 75% of your viewers are saying a phrase that you filtered out. And now because you have an automatic delete function, now your comments have dropped 80% or something like that. So it probably would be cool to have statistics for stuff like that."* - P9

Table 2: Design goals identified in our needs analysis, and the Findings subsection from which each design goal arose.

#	Design Goal	Connection to Findings
G1	Incorporate a greater degree of organization	3.3.1
G2	Include spelling variants of configured phrases	3.3.2
G3	Preview expected effects while configuring word filters	3.3.3
G4	Offer importing of word filter categories	3.3.4
G5	Show the effects of configured word filters	3.3.5
G6	Allow sharing of word filters with other creators	3.3.6

3.3.6 Sharing word filters. Participants expressed diverging views on whether and how they would like to share their word filters with others. Four creators noted that they already use lists configured by other creators by manually copying and pasting them, but found doing so inefficient. Five participants also had a desire to collaboratively create word filters with other creators with similar interests.

> *"A collaborative list would be nice, especially for female YouTubers because we have the most hateful comments, so there's some times that someone receives a bad comment with a word that another didn't have yet, but can appear sometimes, so collaboratively creating these lists would be great."* - P14

Two participants noted that they configure private or sensitive information such as their residential address in word filters to deter posters from doxxing them. Therefore, they would want to keep certain categories of word filters private. Three creators worried that allowing everyone to share their word filters with one another would make them open to exploitation by bad actors or may proliferate poorly configured word filters that result in greater inaccuracy. P15 said:

> *"If I know a particular creator,...and I already buy in and trust the way that they run their community and the way that they keep people safe within their community, then I am more likely to say, I know where this came from, I know there's likely reasoning behind most of it, and someone has been thoughtful about this...If it's just everybody always contributing, while that's probably easier to write new lists, it's probably not thoughtful enough to keep the gates closed to people who would abuse it."* - P15

Summary. We determined several design goals necessary for improving the usability of word filters. First, participants noted that the current user interface for blocked words is a simple text box on most platforms, which does not meet their need for a greater degree of organization. They wanted the ability to configure multiple categories of word filters so that they could view and moderate different types of comments differently (G1). Another issue with a comma-separated list of phrases is lack of organization around spelling variants of a phrase, which participants expressed that they often wanted to include (G2). Third, they wanted more visibility into the functioning of word filters, especially through observing examples of comments that would be caught by the filters they configure (G3). Participants also wanted the ability to import existing categories of word filters to reduce the labor of initial configurations (G4). They were eager to see statistics about word filters' actions over time (G5). Finally, some participants wanted to share their word filters with other creators (G6). Table 2 summarizes these design goals.

4 FILTERBUDDY: A WORD FILTER TOOL FOR CREATOR-LED MODERATION

As described in the previous section, our needs analysis enabled us to determine a set of design goals that would empower online content creators. With these goals in mind, we next designed and built *FilterBuddy*, a system for creators to better configure word filters. We developed FilterBuddy for YouTube, the platform of choice for many creators, after we observed that it contains only basic tools for configuring blocked words. However, FilterBuddy is a general framework applicable to any content sharing platform. We next describe a scenario inspired by our subjects of how FilterBuddy can be used, followed by features and implementation of the system.

4.1 User Scenario

Isabella is a YouTube creator who likes sharing videos of herself cooking authentic Iranian recipes on her channel. She recently posted a video where she shared her spiced meat kebab recipe, which went viral and as a result, her channel suddenly receives thousands of new subscribers and comments. While Isabella appreciates many of the encouraging comments and requests for sharing other Iranian recipes, she also receives a few comments that criticizes her looks and her clothing. Some other comments criticize her Iranian heritage while a few others are blatantly Islamophobic. Isabella likes to keep her channel conversations focused on celebrating Iranian food; therefore, she manually starts removing inappropriate comments. However, she quickly realizes that this manual process would be very time consuming. She also notices that many comments include the same or similar offensive words.

Looking to reduce her stress and time investment in removing undesirable comments, she logs in to FilterBuddy using her YouTube credentials. She notices that FilterBuddy already provides some importable categories (Figure 4), and she chooses to import a category for *Misogyny*. She is redirected to her new Misogyny category page (Figure 2) that is pre-configured with 23 misogynistic phrases, each phrase listed alongside the number of comments on her channel containing that phrase (Figure 2-D). She sets 'Delete Comment' for some phrases and 'Send to Review Folder' for other phrases as the action to take on matching comments. Noticing that her channel also contains anti-Islam comments, she creates a new category from scratch called *Islamophobia*. When adding each new phrase to this category (Figures 2-E, 3), FilterBuddy shows her all

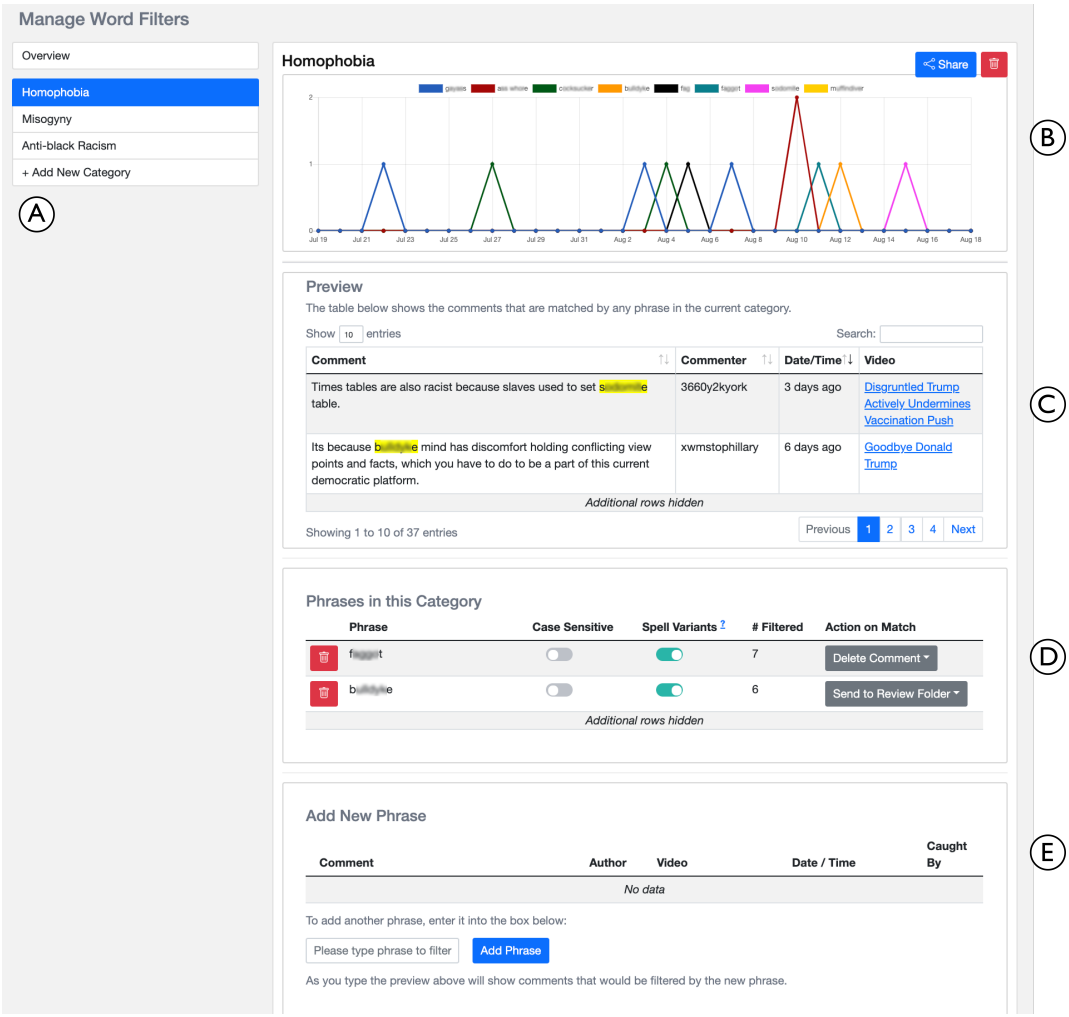


Figure 2: A screenshot of FilterBuddy's category page showing (A) a sidebar with links to the overview page, each configured category page, and *Add new category* page; (B) a chart showing the number of comments caught by each category phrase in the past month; (C) a paginated table previewing all comments caught by the category; (D) a table of phrases configured in the category with options to include/exclude spelling variants and determine action on match for each phrase; and (E) a section to add new phrases in the category. Note that we limit the number of table rows we show in all the figures for brevity.

the prior comments caught by that phrase—reviewing these comments lets her make informed decisions about whether to keep or remove the phrase.

A week later, she logs in to FilterBuddy, and her home page (Figure 5) shows her a graph with the temporal trends of her categories of Misogyny and Islamophobia filters. She is glad to notice that the number of misogynistic comments reduced in the past week. She also scans her most recent comments on this page and notices that some comments not caught by any existing category are politically divisive. Desiring to remove these, she creates a new category for *Politics* to remove comments with political content.

4.2 FilterBuddy Features

We now discuss how the features of FilterBuddy serve the design goals identified in Section 3.3 (Table 2).

4.2.1 Incorporating a greater degree of organization (G1). Instead of organizing around individual keywords like most word filter tools, FilterBuddy organizes around *categories*, where each category contains a list of phrases (see Figure 2-A). This allows for greater organization when managing a large number of word filters. Users can also receive analytics about categories instead of every single phrase, which can help with understanding higher-level trends.

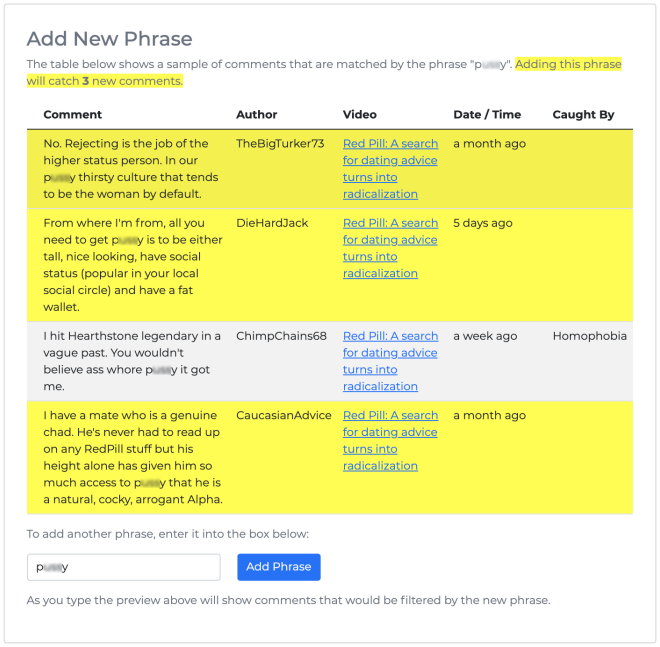


Figure 3: 'Add New Phrase' section on the FilterBuddy Category Page. As the user types in a phrase, the comments caught by that phrase are auto-populated. Comments not already caught by any configured phrases have a yellow background so that they are easily distinguished.

4.2.2 Configuring spelling variants for each phrase (G2). Each phrase in FilterBuddy is not just an exact-match filter but a regular expression under the hood. We provide two forms of per-phrase settings, including whether matching should be case-sensitive and if it should cover spelling variants (see Figure 2-D). For example, 'ABCD' is an acronym that is a South-Asian slur, which, if configured with cases-sensitive setting on, will be caught in comments only where the term occurs with the capitalized case. With the *spell variants* setting turned on, a phrase will also match plurals and words with characters repeated. For example, configuring *'shit'* as a phrase would also capture comments containing *'shiiit'* and *'shittt'*. In the future, we will investigate more comprehensive approaches to capturing spelling variants while retaining some customizability. By tying these settings to a single phrase, we also increase organization (*G1*), as typically, users would need to write out every variant as distinct comma-separated keywords. FilterBuddy allows configuring the same phrase as a word filter in more than one category: when a word filter configuration is changed in one category, the same change also reflects for that filter in other categories.

4.2.3 Previewing comments caught by a phrase (G3). FilterBuddy has a dedicated section '*Add New Phrase*' (Figure 2-E) in each category page for users to add new phrases. As the user types in a new phrase, FilterBuddy will interactively show a preview of all previously published comments that would be matched by that phrase (see Figure 3). This preview also shows whether these comments have not already been caught by another phrase and reports the total number of uncaught comments at the top. We note that this feature may inadvertently expose users to abusive comments they receive. However, by letting users understand the contextual details of how particular keywords are used on their channel, it can help them make more informed decisions on how to automate moderation and see fewer of these comments in the future. We expect that creators will need to review only a few comments to make such decisions and once set up, more accurate configurations will better serve creators' needs (see Sec. 3.2.3) over a longer duration.

4.2.4 Offering importable categories (G4, G6). Users can click '*+ Add New Category*' in the FilterBuddy sidebar (Figure 2-A) to create a new category. This redirects them to a page (Figure 4) where they can name and create an empty category. FilterBuddy also offers a set of built-in categories to make it easier for users to get started with creating word filters. Currently, FilterBuddy has the following importable categories: *Homophobia*; *Physical Violence*; *Sexual Violence*; *Pejorative Terms for Women*; and *Anti-Black Racism*. We manually collected these lists of terms from searching through academic and other public resources. This selection of categories was especially aimed at addressing the identity-based attacks that our female, LBGTQ, and black participants reported in our needs analysis. For each of these built-in categories, we show an information box containing the following information:

(1) A brief description of the category
(2) The authors (website, organization, or researchers) that developed the category
(3) The number of rules contained in the category
(4) Examples of rules in the category

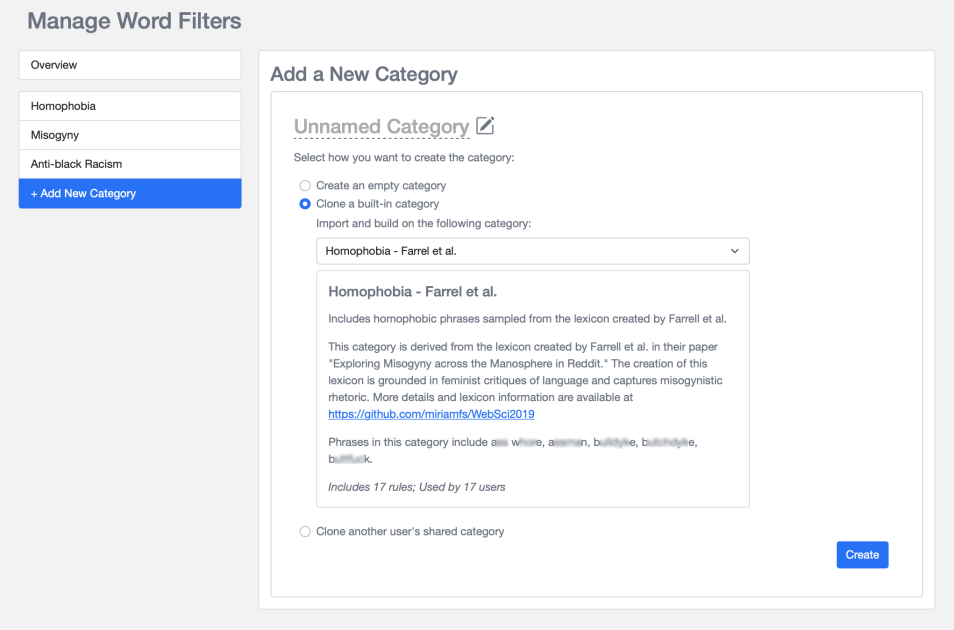


Figure 4: FilterBuddy's 'Add a New Category' Page. Users can either create an empty category, import one of the built-in categories, or clone a category shared by another user to quickly set up their configurations. We show here the details of the built-in 'Homophobia' category selected in the dropdown.

(5) The number of users who have imported this category

After importing a category, users can customize their copy of the category just as if they made it from scratch.

Finally, on this page, users can also import a category that has been shared with them by another user on the FilterBuddy site (*G6*). For the purpose of our user study, we pre-loaded the dropdown list with categories that we had authored as a demonstration of this possibility; in a deployment in the wild, the interface would be populated by categories made by friends of the user. We leave to future research a deeper exploration of how to design sharing features such that users can choose who to share each category with (either a selected group of friends or everyone) and how they can use that category (importing versus collaborating). We see file sharing permissions on sites like Google Drive and collaboration models like Git as promising models for designing FilterBuddy's sharing features.

4.2.5 Showing a time-series graph and table of comments (G5). On its home page (Figure 5), FilterBuddy offers an overview of how different categories are functioning recently. This includes a graph that shows the number of comments caught by each category over the past thirty days. It also shows a table with all comments posted on the user's channel ordered by recency, along with sorting and searching capabilities. The table shows for each comment the category, if any, that caught the comment. Combined, these features help users quickly understand how their word filters are operating. On the category page, FilterBuddy presents a similar graph (Figure 2-B) with the phrases that caught the most comments from the category. There is also a table (Figure 2-C) for each category that shows all comments caught by that category, with the relevant phrases highlighted.

4.3 System Implementation

FilterBuddy is a Django web application. It uses a MySQL database to store user data. Users log into FilterBuddy using their YouTube credentials. After they log in for the first time, FilterBuddy uses the YouTube API to retrieve all public comments posted on their channel and saves it to the database. As new comments get posted on YouTube, FilterBuddy periodically checks for and retrieves new comments. Next, depending on how users configure their word filters, FilterBuddy takes actions on new comments as they are retrieved, such as removing caught comments, blocking users, etc., again using the YouTube API. These actions are reflected not just in FilterBuddy's local database but also on the creator's YouTube channel; FilterBuddy is not just a simulation but a tool designed to be actively used by creators. FilterBuddy does not currently support bulk or retroactive actions on older comments though this may be incorporated in the future. We have released FilterBuddy's code on Github[2] as open source, and we plan to publicly launch FilterBuddy as a free hosted service in 2022.

5 EVALUATION

We introduced different features of the FilterBuddy tool and its workflow to eight YouTube creators (referred to as U1-U8), and then asked participants to explore the tool and author filters over

[2]https://github.com/Social-Futures-Lab/FilterBuddy

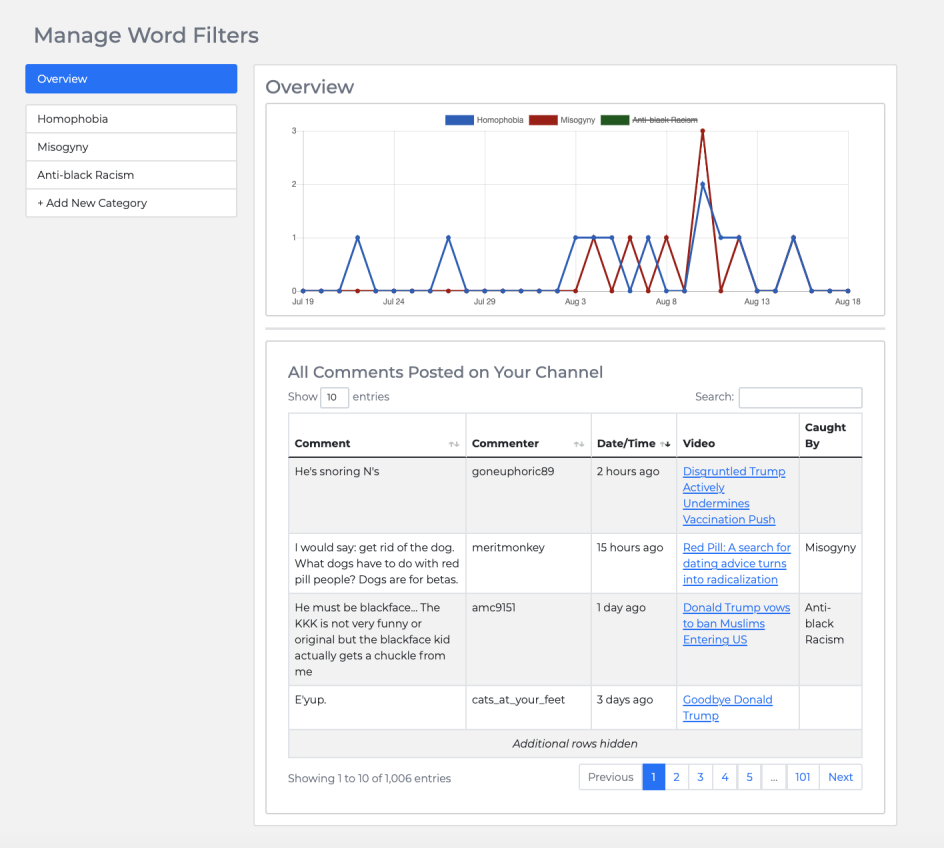


Figure 5: FilterBuddy's Home Page. It shows (a) a time-series graph for number of caught comments aggregated at the category level and (b) a paginated table of all comments posted on the user's channel.

their comments while providing qualitative feedback. In this section, we discuss our methods and findings from this user study.

5.1 Methods

We reached out to the same participants we interviewed during our needfinding study and also conducted an additional round of social media recruitment. Five participants were previously involved in our earlier study, and we also gained three new participants. Similar to our recruitment criteria in the prior needs analysis (Sec. 3.1.2), we oversampled gender, racial, and sexual minorities. We ensured that our participants had received at least dozens of comments on their videos, and they were familiar with settings available to automate content moderation. Table 3 details the demographic information about our participants and the size and topical focus of their channels. Each session was conducted online via Zoom for 45-90 minutes, recorded, and transcribed. Before each session, we sent the participant details of our project and offered to answer any questions they had. All participants again received $25 as compensation for their participation.

We began each user study session by providing a brief overview and goals of FilterBuddy. Next, we asked the participant to log into FilterBuddy using their YouTube credentials. During this step, FilterBuddy automatically retrieved all public comments posted on the participant's channel and populated the FilterBuddy interface. We then asked participants to explore the interface and create new word filter categories, either starting from scratch or importing a built-in or shared FilterBuddy category.

After authoring word filters, the FilterBuddy graphs and tables become populated with the comments posted on the participant's channel as if those filters had been turned on for the last thirty days. This was necessary to demonstrate the tool within the time period of a user study session; otherwise, participants would need to turn on the tool for an extended period of time. We chose in this study to not actually carry out any actions arising from any of the word filters that participants created, so as to minimize disruption to our participants and encourage freer exploration. For the study, the tool only retrieved comments from participants' channels that were publicly readable, and we deleted any user data collected by the tool after their user study session was over.

We asked participants to "think-aloud" [65, 116] as they tried different FilterBuddy features, expressing their likes and dislikes and describing additional features they would like to have. We asked them to reflect on how they would configure FilterBuddy if they were actually using it to moderate their channels and imagine

Table 3: Demographic information of user study participants.

Sr. no.	Age	Gender	Country	Platforms Used	Previously	Size	Topic
U1	23	Male	USA	YouTube, Twitch	P13	N/A	Gaming
U2	20	Female	USA	YouTube, TikTok	P4	<500	Music creation
U3	24	Female	Turkey	YouTube, Twitch	-	<500	Gaming
U4	24	Female	USA	YouTube, Twitter, Twitch	-	>100k	Makeup
U5	29	Female	Germany	YouTube	-	10k-20k	Academia
U6	24	Female	USA	YouTube	P3	20k-50k	School life
U7	31	Female	Germany	YouTube, Instagram	P14	10k-20k	Science
U8	29	Female	USA	YouTube, Twitch	P12	5k-10k	Miscellaneous

themselves coming back to this tool after having it on for thirty days.

We again conducted interpretive qualitative analysis [78] on the transcripts of the user study sessions. We also referred to the code book we had used for our needs analysis since it helped us associate our design goals with participants' evaluation of features designed to address those goals. We now present the prominent themes we observed during this process.

5.2 Findings

5.2.1 Participants emphasized control over greater automation. All participants appreciated that FilterBuddy takes care of some spelling variants of phrases and had ideas for additional variants they wanted FilterBuddy to capture. However, beyond common heuristics such as substituting 'e' with '3', most participants were wary about more sophisticated tools for automatically learning what content to filter. Instead, all study participants expressed a desire to retain control over and visibility into the operations of any moderation tool they adopt. U3 expressed concern about the accuracy of machine learning tools and wanted to see their execution and results before adopting them. In a similar vein, U6 said:

> *"I would prefer to just do it myself manually. I almost wouldn't want that power taken away from me because moderation is so personal and difficult and very sensitive, so I think I would want complete control over it." - U6*

However, participants had ideas of ways that automation could be introduced but without a loss of control. For instance, four participants wanted automated suggestions of new phrases to add to their existing categories, with the ability to choose whether to incorporate them.

5.2.2 Importable categories were seen as a powerful way to reduce toxic content. Participants noted that it can be challenging to spontaneously think of phrases to configure, and they often learn about candidate phrases to block only after they see them appear in their comment feed. U6 also found it discomforting to explicitly type out offensive phrases like the N word. Therefore, most participants listed the ability to import built-in categories as their favorite feature of FilterBuddy. For example, U4 said:

> *"Having the base list is just perfect for the creator, because they don't even have to like spend the time, just like brewing over what words might be used towards them, which is like the most messed up thing to have to think about in the first place, so having a tool like this is just amazing. Honestly, it's a lifesaver!" - U4*

Five participants noted the potential of using built-in FilterBuddy categories to broadly improve the content quality on platforms like YouTube and desired to see their widespread adoption. U8 suggested a new FilterBuddy feature that lets users propose additional phrases to include in built-in categories, so that they continue to respond to the evolving needs of marginalized users. U8 also recommended that FilterBuddy encourage creators to import categories such as racism and homophobia to build more inclusive channels:

> *"The creator is a cultural lead and, if you allow stuff that, oh it doesn't apply to me but like it's still gross, you're kind of defining that you allow that gross behavior, so some little reminder of like 'even if you're not black, consider using this category' would be great." - U8*

5.2.3 Participants wanted defense mechanisms to prevent false positives. Our participants showed a keen awareness of how overly-broad FilterBuddy configurations can create false positives and drive away users. U6 expressed concern that spelling variants of phrases might unintentionally include some acceptable phrases and suggested that users be shown all English dictionary terms that the spelling variants include and be able to make exceptions for those terms. U3, who runs a Turkish YouTube channel, worried that configuring some English terms might unintentionally catch some acceptable Turkish terms; therefore, she appreciated that the system lets users toggle on/off the spelling variants of phrases. U5 wanted to build more complex rules that filter comments containing multiple specified phrases.

Participants appreciated the analytics and visualization features of FilterBuddy that helped them reduce false positives. They liked the ability to scan through and edit the specifics of all phrases and their corresponding actions in each category. All participants noted that seeing a preview of caught comments when adding a new phrase helped them make more informed decisions about setting up their filters:

> *"I like that you can kind of check for how bad something might be here before you add it, I think that's pretty cool!" - U8*

> *"I think that this is actually very helpful because it helps you understand how many of your word choices are incredibly common. For instance, if you were to use*

> *a word that's a super common phrase and this table shows that a lot of people are going to end up using it generally in any comment, that helps me understand like how much I should narrow it down compared to what I was doing." - U4*

Participants also valued seeing other metadata about caught comments. For instance, U6 considered adding a swear word as a phrase, but, when seeing that the preview of caught comments for that word included amusing comments posted by her friend, she decided against doing so. Four participants noted and appreciated the visual distinction of seeing which comments were not already caught by existing phrases.

Importing built-in FilterBuddy categories or those developed and shared by other creators was seen as a way for false positives to creep in. Participants noted that they would like to personalize and maintain their own copy of the imported categories. If new phrases were subsequently added to the original categories, they wanted them to appear as suggestions to add on their category page rather than having them automatically added:

> *"My ideal would be like, 'Hey, changes were detected to this one that you copied from. Do you want to keep those, like, do you want to merge the changes in to your list - yes or no'." - U8*

This focus on reducing false positives is in line with creators' desire to maintain a balanced approach to content moderation we discussed in section 3.2.3. Indeed, we observed that creators who wanted to grow their channel's subscriptions were more wary of having user comments incorrectly removed. This is in contrast to Reddit moderators' focus on minimizing false negatives as opposed to false positives when using Reddit Automod, as observed by Jhaver et al. [53]. One explanation for this contrast could be that Jhaver et al. [53] studied large Reddit communities whose moderators were more concerned with reducing inappropriate comments than growing their subreddits.

5.2.4 Participants wanted to use FilterBuddy for more than just removing offensive speech. While every participant wanted to set up categories to address identity attacks, such as sexism, racial slurs, homophobia, and Islamophobia, other use cases were also frequently brought up. U4 and U8 said that they would create a category containing phrases related to their personal information, such as social security number and address; they intended to use this category to preemptively prevent doxxing. U6, a female Asian creator, expressed a desire to set up a category to filter out comments that remark on her appearance using phrases such as 'ugly', 'fat', and 'unattractive.' Five participants also discussed creating categories to remove off-topic, especially politically charged comments.

All participants noted the utility of FilterBuddy as a tool to filter comments in different buckets and better understand their audiences. Four participants wanted to set up a category to capture comments with positive emotions but not perform any action on them. For instance, U7 wanted to set up categories for positive and negative phrases and visualize how the volumes of comments caught by both categories change over time. Participants also shared a number of creative categories and use cases that we had not anticipated. For example, U4, who operates a gaming channel, said:

> *"I would probably make a self-promotion category because I know a lot of times you'll have people being like, 'Oh, subscribe to my channel instead' or whatever, because it's not quite a hate comment, it's just something that's kind of frowned upon. Critical or constructive criticism comments would probably have like its own separate category. Maybe like a collab category so if anyone wanted to say, 'Oh, you should play with me sometimes,' something like that, then it just filters those. Maybe game recommendation because I definitely play a lot of games, so say a phrase like 'You should play…', I'd put that in there; it gives you an idea of how often are people saying I should play something else." - U4*

5.2.5 Participants were keen to engage in community building with other creators by leveraging the sharing feature. Participants reacted positively to the idea of letting other creators benefit from their word filters, especially when context-specific moderation is required, and multiple creators have similar moderation needs. For example, U3 noted that FilterBuddy's built-in categories might not cover all moderation requirements; therefore, letting users share categories may fill essential gaps:

> *"Sharing is really useful because you will not be able to think of everything that is a problem in the Internet. So in Turkey, there may be things that are not a problem in America or in Korea or Europe. So, if a Turkish YouTuber designed her categories to be exactly what other Turkish YouTubers would want, it will be really awesome to be able to share her categories and will be really time saving, too." - U3*

U8 wanted the ability to share not just the phrases she set up in any category, but also the actions configured for each phrase, and information about whether she turned on the case-sensitive and spelling variant setting for that phrase:

> *"I think, as long as they know they can change things themselves, it is not a bad idea to have it shared the way that creator set it up because if I were passing it on to somebody else, I want to make it as seamless as possible…." - U8*

5.2.6 FilterBuddy was seen as a powerful tool to help content creators. Overall, participants were excited to use FilterBuddy and appreciated its utility in not just removing unwanted content but also in helping them understand topical trends in their comments. Six participants appreciated the time-series charts for different categories and three of them expressed an interest in using the graph spikes as clues to further filter and investigate comments causing those spikes. U6 predicted that FilterBuddy would reduce the emotional labor of removing inappropriate comments. U8 anticipated that FilterBuddy would deliver substantial time savings for creators:

> *"I think it is empowering and also good for people's growth because, whether you're a big YouTuber with you know, three editors and a manager or if you're a small creator who just makes videos in their free time, that's still time you don't have to spend on manual moderation." - U8*

Participants also commented on benefits for marginalized groups since such groups receive disproportionate number of undesirable comments. For instance, U8 felt that black and female content creators would especially benefit from using this tool because of its proactive support to address sexism and racism. U2 noted that creators who have been targeted by commenters in the past would be especially motivated to use this tool to preemptively and automatically address online attacks. Overall, participants did not find it cognitively challenging to understand how FilterBuddy works and each participant appreciated the confluence of multiple useful features that it offers.

> *"If you want something generally family friendly, this is just like a tool to help that happen. It's a one stop shop to solve moderation issues and manage your comments."*
> *- U1*

6 DISCUSSION

6.1 Design Implications

Drawing on both our system evaluation (Sec. 5) and needfinding results (Sec. 3), we present the following design implications that navigate different trade-offs we identified between creator goals.

6.1.1 User control when using automation. One of our key findings from both our needfinding (Sec. 3.3.3, 3.3.5) and system evaluation (Sec. 5.2.1) is the extent to which content creators value retaining control and having visibility over their moderation operations. As noted earlier, study participants preferred to retain granular control over the actions associated with each configured keyword and to specify whether they want to include spelling variants (Sec. 3.3.2). However, additional control can incur a trade-off with manual effort. Turning on the spelling variants setting has the advantage of reducing the number of separate rules that creators would need to set up for terms with similar meanings. On the other hand, turning this on also reduces visibility into exactly which keywords are currently set up to trigger the configured actions and increases the chance of obtaining false positives. Participants were also wary of catching false positives with greater automation, which made them hesitant to use machine learning tools to remove inappropriate content (Sec. 5.2.1). This highlights the importance of reducing burden on users with sensible defaults in place that still let them retain granular oversight.

Prior research has surfaced similar tradeoffs in other domains involving human-computer integration [31] and explored questions such as the type of tasks that should be delegated to AI, the humans' preferences for what role AI should play, and ways to evaluate the division of labor between humans and automated systems [2, 33, 68, 70]. Evaluating a dataset of 100 human-machine collaboration tasks, Lubars and Tan found a disinclination towards "AI-only" designs and a preference for machine-in-the-loop designs [68], a finding echoed in our study. They also found trust in AI system to be an important factor in shaping human preferences of optimal human-machine delegation—this is in line with our participants' preferences for easily controllable, transparent regular-expression based moderation tools. Mackeprang et al. highlight the challenges of finding the right-level of algorithmic support that achieves the compromise between minimizing human effort and maximizing system performance [70]. Their methods of systematically defining and evaluating different levels of automation could serve as a template for further development of AI tools for content creators.

We see more advanced AI approaches playing a critical role in further strengthening the value of tools like FilterBuddy. For example, training data based on the labels generated by configured word filters can be used to train ML models to obtain phrase suggestions for users to add to their existing categories. Further, collaborative filtering techniques [100] can be employed to suggest creators additional categories to configure. However, for any ML additions to FilterBuddy, developers should remain mindful of the creators' need to maintain agency over their tools. More broadly, we carry forward the arguments made elsewhere that when integrating ML in the workflow of human-managed systems, designers and developers should carefully attend to the needs of their users that go beyond traditional metrics, such as precision and recall of ML models [24].

6.1.2 Designing features to better understand and be responsive to audiences. The study participants' enthusiasm about developing custom categories (Sec. 5.2.4) and visualizing resulting graphs (Sec. 5.2.3) is indicative of content creators' need to better understand their audiences. This suggests that creators would benefit from advanced analytic and visualization tools that facilitate discovery of unusual temporal and topical trends in their comments. For example, creators may value receiving notifications of unusual spikes in posting activities or the use of toxic language on their channels. We were also surprised to find that participants wanted to use our tool for more than moderating out unwanted content (Sec. 5.2.4). It indicates that creators may value new, creative ways of understanding their audiences where they have the power to specify which phrases or activity they want to monitor.

As content creators invest substantial amounts of time and emotional labor to gather and sustain an online audience [13, 39], we also found that fears over audience backlash and exodus in response to excessive moderation are a key factor in creators' moderation strategies (Sec. 3.2.3). While the configuration assistance and performance visualizations provided by FilterBuddy are helpful in setting up rule-based moderation (Sec. 5.2.3) that reasonably balance these goals, creators still need to exert significant care in their decisions about which filters to configure. Tools could additionally incorporate features such as ways for creators to have a trial run of a new filter or be able to revert a filter that was instated and bulk restore comments that were removed.

6.1.3 Sharing and collaborating on word filters. Study participants expressed an interest in sharing word filters with one another to let other creators benefit from their configurations (Sec. 3.3.6, 5.2.5). However, they were also concerned that sharing these word filters might raise the possibility that these configurations are leaked, rendering them vulnerable to exploitation by bad-faith actors who use creative misspellings to bypass word filters (Sec. 3.3.6). Tools will need to consider how users can signify trust to each other when deciding whether to share a word filter. However, it may be the case that even a publicly available word filter still has utility as only few bad actors are motivated enough to seek to bypass a filter. This is our rationale for including built-in word filters gathered from public resources.

Similarly, participants expressed an interest in collaboratively co-creating word filters (Sec. 3.3.6), but poor configuration changes by one creator could disrupt the moderation for all collaborators. In addition, our tool currently supports only the cloning of existing word filters, as opposed to a subscription or forking model, where changes to the original filter can flow to subscribers. Also, we note that many creators have a small team of volunteer or paid moderators who help them with reviewing comments; our tool at the moment only has a login for the creator of the channel. In the future, collective governance mechanisms will become increasingly important as our tool moves from being primarily used individually to author one's own filters towards more collective ways of moderating together or for each other.

6.2 Role of Different Stakeholders

We found in our needs analysis that much of the burden of content moderation rests on just one stakeholder—the content creator. This is especially challenging for creators who belong to minoritized communities and receive inordinately high volumes of inappropriate comments. Previous research has examined how viewers can play a crucial role in supporting such creators by volunteering to serve as channel moderators, helping creators configure appropriate word filters, and endorsing through donations [71, 105, 124]. As demonstrated by our tool and prior research, creators themselves can also help one another by co-constructing shared resources including word filters, starting a dialogue on moderation strategies, and engaging in collective action to demand better working conditions [19, 81]. Drawing on our design exploration with FilterBuddy, we reflect here on how other key stakeholders, including platforms, third-party organizations and advocacy groups, and policymakers can alleviate some of the pressures and assist creators in flourishing healthy online spaces.

6.2.1 Platforms. From our examination of existing word filter tools on major platforms, we found that they lack many basic usability features for organization, visualization, and interactivity, and these omissions lead to frustration for content creators. Our subsequent development and evaluation of FilterBuddy confirms that these features are both readily implementable using standard techniques and highly desirable to creators. The question then remains why major platforms, who have considerably more resources than we do to design and develop such features, have not done so yet. We call upon platforms to dedicate more resources towards building more powerful creator-led tools for moderation. Alternatively, platforms could go beyond providing API access to their data and dedicate technical and human resources to foster an ecosystem of third-party tools such as FilterBuddy. They can demonstrate their commitment to improving creators' work conditions by collaborating with academic researchers and minority support groups to build long-needed moderation tools [72].

6.2.2 Third-party organizations and advocacy groups. As the response to FilterBuddy shows, researchers, volunteer users, and third-party organizations can play a critical role in supporting creator communities by devising and deploying novel moderation solutions. These groups can engage closely with marginalized content creators, understand their evolving moderation and information needs, surface them to platforms, policymakers, end-users with computational expertise and the research community, and contribute resources to address those needs.

Our participants highlighted the ability to import built-in platform categories as their favorite feature of FilterBuddy. This emphasizes the importance of offering technical resources—such as carefully curated lexicons—in addressing online hate. Recent work has shown how widely-adopted word filter lists such as the List of Dirty, Naughty, Obscene and Otherwise Bad Words (LDNOOBW) can harm marginalized groups, such as by censoring terms related to LGBTQ topics, due to the lack of input from members of those groups [25]. There is an opportunity here to involve minority support groups and use their domain expertise and influence to curate and publicize appropriate lexicons. By creating such lists and making them easy to adopt and fork in FilterBuddy, advocacy groups could promote social norms that discourage the use of toxic language such as racist slurs and homophobic swear words.

6.2.3 Policymakers. While prior HCI policy research has largely focused on the forms of public law that sweepingly regulate technology design, it is important to recognize that formal law intersects with procedures set by private firms and individual regulators in complex and nuanced ways [12, 52, 60]. For example, formal law on what is considered as hate speech would likely shape the word filter categories adopted by FilterBuddy users. It is, therefore, vital that policymakers attend to how public policy shapes technical design and social practices around moderation tools for creators.

One crucial way to empower creators is for policymakers to mandate that larger platforms must provide creators with moderation mechanisms that let them efficiently manage their channel's comments. Such mechanisms may include the ability to author word filters, configure spelling variants, and examine their performance over time. It has been well recognized that legislative focus on certain protected characteristics such as race, gender and religion has inadvertently excluded other vulnerable groups such as immigrants [113]; policies to mandate authoring of personalized filters can help protect such groups. Policymakers can also incentivize platforms to enable APIs, plugins, or protocols to create a third-party marketplace of governance services [30, 74, 103, 126], that will further empower content creators.

6.3 Ethical Considerations

Power of moderating content online can be exploited to censor important conversations [37, 84, 110, 122]. Rule-based moderation tools like FilterBuddy that allow auto-removal of comments containing certain keywords can further facilitate this process. Deploying such tools to reduce harm, therefore, raise enduring, familiar tensions between libertarian tendencies (e.g., freedom of speech) and authoritarian practices (e.g., censorship) [28, 113, 119–121]. However, we note that the ability to block configured keywords already exists on most platforms, including YouTube, in the form of word filters. In fact, FilterBuddy offers many analytic and visual aids to avoid setting up configurations that result in false positives, so that well-meaning creators do not inadvertently censor valuable comments. Yet, third-party moderation tools (e.g., Twitter blocklists

[35, 57]) have previously been exploited in ways unanticipated by their creators. Therefore, we are committed to conducting a regular oversight of who uses FilterBuddy and how it is being used.

7 LIMITATIONS AND FUTURE WORK

Going forward, we plan to implement additional features that study participants requested in our evaluation (Section 5). This includes incorporating additional types of spelling variants, creating a time selector and video selector for graphs, and configuring more complex word filters that identify, for example, the co-occurrence of multiple phrases in the same comment. After additional development to improve security, we will release FilterBuddy as a public hosted tool.

Currently, FilterBuddy is a YouTube-only tool; however, it could be extended to other platforms that support creator-oriented communities with little additional development. The YouTube API features that we use, including comment retrieval and comment removal, have uneven but growing support across other major platforms. Many of FilterBuddy's features are also valuable and much needed in other moderation contexts. For example, prior research has shown that volunteer moderation teams on sites like Reddit often struggle with using moderation tools that are difficult to configure and have lack of visibility into the operations of such tools [53]; FilterBuddy offers to address these needs.

FilterBuddy relies on analyzing sequences of written words to automatically detect undesirable content. This entirely text-based approach is insufficient to regulate online communication that is frequently and increasingly multimodal. Thus, it is important to incorporate and build upon prior research on multimodal approaches to hate speech detection in the future [49, 127]. Further, since FilterBuddy is a regular expression-based tool, it is limited in its ability to detect posts that use more subtle or complex linguistic strategies to propagate hate speech. Similarly, it might have limited utility in addressing some forms of undesirable content such as political or health misinformation. Yet, we expect that this tool will introduce friction [14] to successfully posting inappropriate posts and act as a powerful deterrent against bad actors, thereby delimiting their submission of such posts. Prior research has shown that poorly implemented blocklists can disproportionately remove text from and about minority individuals and exacerbate existing inequalities [25, 91, 98]. We hope that FilterBuddy's analytic and visualization features to configure more accurate word filters would help minimize such harms.

Since we evaluated FilterBuddy using a small user study, we do not yet have empirical evidence about how such a system would work in the wild. While our user study helped us understand how creators would use FilterBuddy and suggested promising design directions, an in-the-wild study would clarify new aspects, such as determining possible incentives towards sharing word filters with other creators, using multiple categories for the same content type, and defending against bad-faith actors. We suspect that our user study may have introduced some biases, e.g., participants may have overstated their desire for control, which may alter as the creators use our tool in the wild and get overworked; we are excited to see if this calculus changes in a long-term field deployment. Even in our small user study, participants devised a number of use cases for creating FilterBuddy categories that we had not anticipated. Therefore, an in-the-wild study would also help us identify further utility of this tool. We are also excited to study the sharing practices among YouTube creators and further strengthen the security and utility of sharing mechanisms FilterBuddy provides.

8 CONCLUSION

In this work, we study the current practices of how online content creators use rule-based moderation tools, finding from 19 interviews that many creators find existing tools cumbersome to use, consider it difficult to come up with keywords to configure, and want analytic information that can guide their configuration and review of these tools. We developed FilterBuddy, a word filter tool that helps content creators better address their moderation needs by offering new visualizations, configuration assistance, and ready-made categories. As opposed to the reactive, too-little-too-late approach primarily used by platforms to decrease online harm, FilterBuddy intercepts potentially problematic posts at an earlier stage of the process, before they have been read by the content creators or their audiences. From a user study, we found that creators appreciate the features that FilterBuddy provides, desire to share their configurations with other creators, and show resistance to replacing the rule-based configurations of this tool with solely ML-based approaches.

ACKNOWLEDGMENTS

We thank the Anti-Defamation League (ADL) for supporting this research through the Belfer Fellowship program. We are also grateful to our reviewers and members of the UW Social Futures Lab for providing exceptionally constructive feedback that helped improve this work.

REFERENCES

[1] Crystal Abidin. 2015. Communicative Intimacies: Influencers and Percieved Interconnectedness. (2015).
[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
[3] Rachel Berryman and Misha Kavka. 2017. 'I Guess A Lot of People See Me as a Big Sister or a Friend': the role of intimacy in the celebrification of beauty vloggers. *Journal of Gender Studies* 26, 3 (2017), 307–320. https://doi.org/10.1080/09589236.2017.1288611 arXiv:https://doi.org/10.1080/09589236.2017.1288611
[4] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment is Perceived as Justified. In *Twelfth International AAAI Conference on Web and Social Media*.
[5] Sean Burch. 2019. YouTube says 'INAPPROPRIATE comments' could result in video demonetization. https://www.thewrap.com/youtube-inappropriate-comments-demonetization/
[6] Jean Burgess and Joshua Green. 2018. *YouTube: Online video and participatory culture*. John Wiley & Sons.
[7] Facebook Business Center. 2021. About Facebook Pages. https://www.facebook.com/business/help/461775097570076?id=939256796236247
[8] Jie Cai and Donghee Yvette Wohn. 2019. What are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *Conference companion publication of the 2019 on computer supported cooperative work and social computing*. 166–170.
[9] Marilyn A Campbell. 2005. Cyber bullying: An old problem in a new guise? *Journal of Psychologists and Counsellors in Schools* 15, 1 (2005), 68–76.
[10] Robyn Caplan and Tarleton Gillespie. 2020. Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media+ Society* 6, 2 (2020), 2056305120936636.
[11] Caitlin Ring Carlson, Luc Cousineau, and Caitlin Ring Carlson. 2020. Are You Sure You Want to View This Community? Exploring the Ethics of Reddit's

Quarantine Practice. *Journal of Media Ethics* 00, 00 (2020), 1–12. https://doi.org/10.1080/23736992.2020.1819285
[12] Alissa Centivany and Bobby Glushko. 2016. " Popcorn Tastes Good" Participatory Policymaking and Reddit's. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1126–1137.
[13] Ngai Keung Chan. 2019. "Becoming an expert in driving for Uber": Uber driver/bloggers' performance of expertise and self-presentation on YouTube. *New Media & Society* 21, 9 (2019), 2048–2067.
[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] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (Dec. 2017), 22 pages. https://doi.org/10.1145/3134666
[16] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting Internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3175–3187.
[17] 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
[18] Clement Chau. 2010. YouTube as a participatory culture. *New directions for youth development* 2010, 128 (2010), 65–74.
[19] A Chen. 2019. How YouTubers plan to take on YouTube for better working conditions. *MIT Technology Review* (2019).
[20] Chloe Condon. 2021. Some of you have never had... https://twitter.com/chloecondon/status/1425197893678366723.
[21] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. 18, 3 (2016), 410–428. https://doi.org/10.1177/1461444814543163
[22] Twitch Creator Dashboard. 2021. Creator Dashboard. https://help.twitch.tv/s/article/creator-dashboard?language=en_US
[23] Stuart Cunningham and David Craig. 2017. Being 'really real'on YouTube: authenticity, community and brand culture in social media entertainment. *Media International Australia* 164, 1 (2017), 71–81.
[24] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. 233–240.
[25] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, Matt Gardner, and Hugging Face. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. (2021).
[26] 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*. 1–13.
[27] Stefanie Duguay, Jean Burgess, and Nicolas Suzor. 2020. Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence* 26, 2 (2020), 237–252.
[28] Ronald Dworkin, Ivan Hare, James Weinstein, et al. 2009. Extreme Speech and Democracy. (2009).
[29] Stine Eckert. 2018. Fighting for recognition: Online abuse of women bloggers in Germany, Switzerland, the United Kingdom, and the United States. *New Media & Society* 20, 4 (2018), 1282–1302.
[30] Jad Esber, Boaz Sender, Ethan Zuckerman, Crystal Lee, Nana Nwachukwu, Oumou Ly, Peter Suber, Primavera De Filippi, Sahar Massachi, Samuel Klein, et al. 2021. A meta-proposal for Twitter's bluesky project. *Available at SSRN 3816729* (2021).
[31] Umer Farooq and Jonathan Grudin. 2016. Human-computer integration. *interactions* 23, 6 (2016), 26–32.
[32] Julia R Fernandez and Jeremy Birnholtz. 2019. " I Don't Want Them to Not Know" Investigating Decisions to Disclose Transgender Identity on Dating Platforms. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.
[33] Jessica L Feuston and Jed R Brubaker. 2021. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
[34] Guo Freeman and Donghee Yvette Wohn. 2020. Streaming your Identity: Navigating the Presentation of Gender and Sexuality through Live Streaming. *Computer Supported Cooperative Work (CSCW)* 29, 6 (2020), 795–825.
[35] 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
[36] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511. https://doi.org/10.1177/1461444818776611 arXiv:https://doi.org/10.1177/1461444818776611
[37] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.
[38] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.
[39] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass.* Eamon Dolan Books.
[40] Michael Green, Ania Bobrowicz, and Chee Siang Ang. 2015. The lesbian, gay, bisexual and transgender community online: discussions of bullying and self-disclosure in YouTube videos. *Behaviour & Information Technology* 34, 7 (2015), 704–712.
[41] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.
[42] William A Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1315–1324.
[43] Elizabeth Fish Hatfield. 2018. (Not) getting paid to do what you love: Gender, social media, and aspirational work.
[44] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing" trolling" in a feminist forum. *The information society* 18, 5 (2002), 371–384.
[45] Zorah Hilvert-Bruce, James T Neill, Max Sjöblom, and Juho Hamari. 2018. Social motivations of live-streaming viewer engagement on Twitch. *Computers in Human Behavior* 84 (2018), 58–67.
[46] Jacob Hoffman-Andrews. 2021. Block Together. https://blocktogether.org
[47] Mattias Holmbom. 2015. The YouTuber: A qualitative study of popular content creators.
[48] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 316 (oct 2021), 24 pages. https://doi.org/10.1145/3476057
[49] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909* (2015).
[50] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: Computationally-derived social signals from account histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[51] Jelani Ince, Fabio Rojas, and Clayton A Davis. 2017. The social media response to Black Lives Matter: How Twitter users interact with Black Lives Matter through hashtag use. *Ethnic and racial studies* 40, 11 (2017), 1814–1830.
[52] Steven J Jackson, Tarleton Gillespie, and Sandy Payette. 2014. The policy knot: Re-integrating policy, practice and design in CSCW studies of social computing. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 588–602.
[53] 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
[54] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 381 (oct 2021), 30 pages. https://doi.org/10.1145/3479525
[55] 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
[56] Shagun Jhaver, Seth Frey, and Amy Zhang. 2022. Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms. *arXiv preprint arXiv:2108.12529* (2022).
[57] 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
[58] Jialun "Aaron" Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), Article 55. https://doi.org/10.1145/3359157
[59] 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.
[60] David Kaye. 2019. *Speech police: The global struggle to govern the Internet*. Columbia Global Reports.
[61] Moira Kenney. 2001. *Mapping gay LA: The intersection of place and politics*. Temple University Press.
[62] Jina Kim, Kunwoo Bae, Eunil Park, and Angel P del Pobil. 2019. Who will Subscribe to My Streaming Channel? The Case of Twitch. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 247–251.
[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*. 1–12.
[64] Patricia G Lange. 2007. Commenting on comments: Investigating responses to antagonism on YouTube. In *Society for Applied anthropology conference*, Vol. 31. 163–190.
[65] Clayton Lewis. 1982. *Using the" thinking-aloud" method in cognitive interface design*. IBM TJ Watson Research Center Yorktown Heights, NY.
[66] Rebecca Lewis, Alice E Marwick, and William Clyde Partin. 2021. "We Dissect Stupidity and Respond to It": Response Videos and Networked Harassment on YouTube. *American Behavioral Scientist* 65, 5 (2021), 735–756.
[67] Jie Li, Xinning Gui, Yubo Kou, and Yukun Li. 2019. Live streaming as co-performance: Dynamics between center and periphery in theatrical engagement. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–22.
[68] Brian Lubars and Chenhao Tan. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *Advances in Neural Information Processing Systems* 32 (2019).
[69] May O Lwin, Benjamin Li, and Rebecca P Ang. 2012. Stop bugging me: An examination of adolescents' protection behavior against online harassment. *Journal of adolescence* 35, 1 (2012), 31–41.
[70] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the Sweet Spot of Human-Computer Configurations: A Case Study in Information Extraction. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[71] Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 586, 13 pages. https://doi.org/10.1145/3173574.3174160
[72] 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.
[73] Alice Marwick. 2015. You may know me from YouTube. *A companion to celebrity* 333 (2015).
[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] J Nathan Matias. 2016. Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 1138–1151.
[76] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. [n.d.]. Reporting, reviewing, and responding to harassment on Twitter. ([n. d.]).
[77] Nathan J. Matias. 2016. The Civic Labor of Online Moderators. In *Internet Politics and Policy conference* (Oxford, United Kingdom). Oxford, United Kingdom.
[78] Sharan B Merriam. 2002. Introduction to Qualitative Research. *Qualitative research in practice: Examples for discussion and analysis* 1 (2002).
[79] Channel Moderation. 2021. How to Use AutoMod. https://help.twitch.tv/s/article/how-to-use-automod?language=en_US
[80] Casey Newton. 2019. The trauma floor: The secret lives of Facebook moderators in America. *The Verge* 25 (2019), 2019.
[81] Valentin Niebler. 2020. 'YouTubers unite': collective action by YouTube content creators.
[82] Safiya Umoja Noble. 2018. *Algorithms of oppression*. New York University Press.
[83] Fayika Farhat Nova, Md. Rashidujjaman Rifat, Pratyasha Saha, Syed Ishtiaque Ahmed, and Shion Guha. 2018. Silenced Voices: Understanding Sexual Harassment on Anonymous Social Media Among Bangladeshi People. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18)*. Association for Computing Machinery, New York, NY, USA, 209–212. https://doi.org/10.1145/3272973.3274057
[84] Katherine Ognyanova. 2019. In Putin's Russia, information has you: Media control and internet censorship in the Russian Federation. In *Censorship, surveillance, and privacy: Concepts, methodologies, tools, and applications*. IGI Global, 1769–1786.
[85] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1114–1125.
[86] Simon Parkin. 2018. The YouTube stars heading for burnout:"The most fun job imaginable became deeply bleak.". *The Guardian* 8 (2018), 2018.
[87] Block Party. 2021. Block party. https://www.blockpartyapp.com/
[88] 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
[89] Emily Pedersen. 2019. " My Videos are at the Mercy of the YouTube Algorithm": How Content Creators Craft Algorithmic Personas and Perceive the Algorithm that Dictates their Work.
[90] Anthony J Pellicone and June Ahn. 2017. The Game of Performing Play: Understanding streaming as cultural production. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 4863–4874.
[91] David Pinsof and Martie G Haselton. 2017. The effect of the promiscuity stereotype on opposition to gay rights. *PloS one* 12, 7 (2017), e0178534.
[92] Bailey Poland. 2016. *Haters: Harassment, abuse, and violence online*. U of Nebraska Press.
[93] Chand Rajendra-Nicolucci and Ethan Zuckerman. 2021. Top 100: The most popular social media platforms and what they can teach us. https://knightcolumbia.org/blog/top-100-the-most-popular-social-media-platforms-and-what-they-can-teach-us
[94] Tobias Raun. 2018. Capitalizing intimacy: New subcultural forms of micro-celebrity strategies and affective labour on YouTube. *Convergence* 24, 1 (2018), 99–113.
[95] Elissa M Redmiles, Jessica Bodford, and Lindsay Blackwell. 2019. "I just want to feel safe": A Diary Study of Safety Perceptions on Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 405–416.
[96] Kathryn E Ringland, Christine T Wolf, Lynn Dombrowski, and Gillian R Hayes. 2015. Making" Safe" Community-Centered Practices in a Virtual World Dedicated to Children with Autism. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1788–1800.
[97] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.
[98] Jonathan Rosa. 2019. *Looking like a language, sounding like a race*. Oxf Studies in Anthropology of.
[99] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.
[100] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*. Springer, 291–324.
[101] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.
[102] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (2021), 1–33.
[103] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Tan, and Amy Zhang. 2021. Modular Politics: Toward a Governance Layer for Online Communities. *Proc. ACM Hum.-Comput. Interact.* CSCW (Oct. 2021).
[104] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277
[105] 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.
[106] Scott Simon and Emma Bowman. 2019. Propaganda, hate speech, violence: The working lives of Facebook's content moderators. *NPR Technology* (2019).
[107] Peter K Smith, Jess Mahdavi, Manuel Carvalho, and Neil Tippett. 2006. An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying. *Research Brief No. RBX03-06. London: DfES* (2006).
[108] 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
[109] YouTube Studio. 2021. Navigate Youtube studio - YouTube help. https://support.google.com/youtube/answer/7548152?hl=en
[110] 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.

[111] Alexandra To, Wenxia Sweeney, Jessica Hammer, and Geoff Kaufman. 2020. "They Just Don't Get It": Towards Social Technologies for Coping with Interpersonal Racism. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 024 (May 2020), 29 pages. https://doi.org/10.1145/3392828

[112] Brendesha M Tynes, Henry A Willis, Ashley M Stewart, and Matthew W Hamilton. 2019. Race-related traumatic events online and mental health among adolescents of color. *Journal of Adolescent Health* 65, 3 (2019), 371–377.

[113] Stefanie Ullmann and Marcus Tomalin. 2020. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology* 22, 1 (2020), 69–80.

[114] Jirassaya Uttarapong, Jie Cai, and Donghee Yvette Wohn. 2021. Harassment Experiences of Women and LGBTQ Live Streamers and How They Handled Negativity. (2021).

[115] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[116] 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).

[117] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '17* (2017). https://doi.org/10.1145/2998181.2998337

[118] Emily A. Vogels. 2021. The state of online harassment. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

[119] Jeremy Waldron. 2012. *The harm in hate speech.* Harvard University Press.

[120] Jeremy Waldron. 2017. The conditions of legitimacy: A response to James Weinstein. *Const. Comment.* 32 (2017), 697.

[121] James Weinstein. 2017. Hate speech bans, democracy, and political legitimacy. *Const. Comment.* 32 (2017), 527.

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

[123] Stephanie N Williams and Annette V Clarke. 2019. How the Desensitization of Police Violence, Stereotyped Language, and Racial Bias Impact Black Communities. *Psychology and Cognitive Sciences–Open Journal* 5, 2 (2019).

[124] 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.

[125] Donghee Yvette Wohn and Guo Freeman. 2020. Audience management practices of live streamers on Twitch. In *ACM International Conference on Interactive Media Experiences*. 106–116.

[126] Stephen Wolfram. 2019. Testifying at the Senate about A.I.-Selected Content on the Internet-Stephen Wolfram Writings. https://writings.stephenwolfram.com/2019/06/testifying-at-the-senate-about-a-i-selected-content-on-the-internet/

[127] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network.. In *IJCAI*, Vol. 16. 3952–3958.