How Do Users Experience Moderation?: A Systematic Literature Review

RENKAI MA, Pennsylvania State University, USA
YUE YOU, Pennsylvania State University, USA
XINNING GUI, Pennsylvania State University, USA
YUBO KOU, Pennsylvania State University, USA

Researchers across various fields have investigated how users experience moderation through different perspectives and methodologies. At present, there is a pressing need of synthesizing and extracting key insights from prior literature to formulate a systematic understanding of what constitutes a moderation experience and to explore how such understanding could further inform moderation-related research and practices. To answer this question, we conducted a systematic literature review (SLR) by analyzing 42 empirical studies related to moderation experiences and published between January 2016 and March 2022. We describe these studies' characteristics and how they characterize users' moderation experiences. We further identify five primary perspectives that prior researchers use to conceptualize moderation experiences. These findings suggest an expansive scope of research interests in understanding moderation experiences and considering moderated users as an important stakeholder group to reflect on current moderation design but also pertain to the dominance of the punitive, solutionist logic in moderation and ample implications for future moderation research, design, and practice.

CCS Concepts: • **Human-centered computing → Collaborative and social computing**

KEYWORDS: Content moderation; moderation; platform governance; literature review

ACM Reference format:

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

1 INTRODUCTION

Moderation has often been framed as a solution to problematic user behaviors online, in which scale, accuracy, effectiveness, and efficiency are at the heart of latest moderation techniques [13,38,118]. However, such framing can be limited in taking into consideration the perspectives and experiences of various stakeholders, such as human moderators, harassers, victims, advertisers, and users impacted by moderation decisions. Human-right activities and international organizations are afraid that content moderation might suppress freedom of expression [94]. Critics from justice perspectives condemn that moderation algorithms mistakenly flag innocuous content and users as spamming while leaving conspiracy theories on Facebook [76]. Media outlets

2573-0142/2023/10 - 278 $15.00

https://doi.org/10.1145/3610069

stress that moderation mechanisms might not enforce content rules with perfect accuracy given people's different linguistic characteristics and cultural backgrounds [93]. Even social media platforms recognize that their moderation algorithms frequently punish users wrongly [6].

In light of these public concerns, there is a growing body of research exploring how users experience moderation, particularly when they receive a moderation penalty such as content removal [44], account suspension [108], or demonetization [66]. However, there is still a lack of systematic understanding of what constitutes a moderation experience, how previous research has framed and investigated it, and what further work needs to be done. Thus, it is an opportune time to conduct a systematic literature review (SLR) to review and synthesize existing moderation literature that focuses on users' moderation experiences. Leveraging the typical strategies that prior exemplary SLRs (e.g., [5,26]) used to gain the initial understanding (e.g., metadata) of prior literature, we put forward our first research question:

RQ1: What are the characteristics of prior work that investigates moderation experiences?

Systematically understanding how prior work reports and shares empirical findings of moderation experiences is important to moderation research. It can shed light on the trajectory of moderation research conceptually and offer design implications for existing moderation practices of online platforms. For example, an early-on research effort in HCI focused on users' rule-breaking behaviors [97] and how community management design, which is oftentimes manually operated by humans, regulates these behaviors for better online communities [63,64]. Then, an SLR on more contemporary moderation literature can supplement a systematic understanding of how users' perceptions and behaviors are shaped by advanced human-algorithms collaborated moderation [45]. Also, users who experience moderation are often found in underprivileged, powerless positions [24,30], so synthesizing empirical findings of moderation experiences can help articulate more transparent, fair, and contestable moderation designs (e.g., [29,109]). Thus, we ask:

RQ2: How does prior work report findings of users' moderation experiences?

Given that researchers from various disciplines ranging from communication to political science have investigated users' moderation experiences, the perspectives they draw from and the aspects of moderation experiences they focus on can be vastly different. Communication and media researchers have largely uncovered users' struggles and lack of user agency under moderation, such as "shadowban," where the visibility of a user is disproportionately decreased [21,85,117]. Legal scholars have tended to measure content policies' unfairness on social media platforms because the platforms scarcely consider the context of user content (e.g., online community culture and norms) [110]. In more recent HCI research, given the sheered volume of user content and that platforms increasingly rely on algorithms (e.g., machine learning) in moderation [38,45], CSCW researchers have found that various groups of end-users, such as gender and sexual minority people [41,109], content creators [66], or players in competitive games [60] experience opaque algorithmic moderation decisions and more users perceive such decisions hard to be re-examined through appeal procedures [67,108,109]. To gain a comprehensive understanding of how researchers across various disciplines conceptualize moderation experiences and to facilitate the cross-pollination of ideas, we proposed a third research question:

RQ3: How does prior work conceptualize users' moderation experiences?

To answer the three primary research questions, we identified and analyzed 42 empirical studies published between February 2016 (i.e., the earliest time of the literature centering moderated users after January 2016, and we detailed this time criterion in Section 3.1) and March 2022 (i.e., the time we started the search), from ten academic databases through Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [71]. We found a

positive research trend that focused on users' moderation experiences in the past few years. These 42 studies have focused on different types of users who experienced a variety of moderation decisions and designs. We further identified a three-stage reaction model that users went through after they experienced moderation decisions, including instant emotions as well as post-moderation perspectives and behaviors. Lastly, we found five primary perspectives that prior researchers used to conceptualize moderation experiences, including perspectives of moderation effect, user agency, ethical values, marginalization, and creative labor.

Based on the findings, we reflect on current moderation technologies and designs through empirical studies that focus on users' moderation experiences and discuss why and how more alternative moderation modes or designs could be implemented besides the typical punitive logic. Before outlining future moderation research agendas, we discuss the need and ways of considering moderated users as stakeholders to inform moderation design, instruct users' behaviors, and build up mutual trust between moderated users and platform.

Our SLR makes the following contributions to the HCI and CSCW communities:

- We present an in-depth review of prior work on users' moderation experiences, which offers a reflective discussion of what existing research has achieved.
- We offer multi-faceted considerations on how moderated users are stakeholders in moderation designs.
- We highlight a variety of future work agendas for moderation research and practices.

2 RELATED WORK

This section foregrounds users' moderation experiences from prior moderation literature, and then points to the necessity of conducting a systematic literature review on moderation literature for the HCI and CSCW communities.

2.1 Understanding Users' Experiences with Moderation

Moderation helps online platforms to be safe places for end-users. Online moderation means governance mechanisms that construct "participation in a community to facilitate cooperation and prevent abuse" [39]. Moderation oftentimes is implemented by the combination of manual and automated procedures. Human workers, e.g., moderators, as hired and oftentimes "invisible" labor [83], practice online moderation for platforms. They flag/label, collect evidence of, or adjudicate users' deviant behaviors, such as hate speech [27,28], adult content, and terrorism speech [72,90]. Given implementing human moderation is costly [39], platforms have also used algorithms to automate certain steps of moderation, such as flagging and adjudications [2,23,45].

Underneath such human-machine coordination in moderation design, multiple stakeholders, such as researchers, designers, and more, play certain roles in making it work but also face challenges. For instance, claiming to be an independent entity of Facebook, the Oversight Board reviews appeals of content moderation cases [119]. However, it neither discloses much rationale for how it adjudicates moderation cases nor is open to all users to discuss its decisions on Facebook [8]. Besides, given that platform rule-makers, as an non-representative part of the end-user population, articulate content rules on platforms [31,110], researchers have criticized that these rules do not sufficiently consider the context of content (e.g., the localized meaning of user content, identities of speakers and audiences) [105,110]. As a result, designers or engineers might face challenges in constructing moderation mechanisms to both equally moderate content and execute the same extent of tolerance or sanction on content based on same levels of harmfulness

[50,86]. Taken together, the challenges faced by these stakeholders of moderation designs are eventually imposed on end-users and reflected by their experiences because they are the ones directly interacting with moderation mechanisms.

Thus, researchers have lately expressed the importance of understanding users' moderation experiences. End-users experience moderation ranging from content removal [44] to deletion of hashtags [14,33] to account suspension [73] to community-wide moderation [15,19]. However, these moderation decisions are often without detailed rationale or explanations attached. Users need to laboriously make sense of or interpret why they experienced moderation [36,84]. Suzor et al. [99] and West [73] shared users' complaints that platforms did not inform what content policies were used for moderation decisions they issued. Sometimes such opaque moderation would bring harm to already marginalized groups. Sybert discovered that gender or sexual minority users condemned the new content policies on Tumblr after they experienced opaque content removal [100]. Besides, pro-eating order (Pro-ED) users felt they lost opportunities for self-recovery and obtaining community support due to content removal [30].

Along with the potential imperfection of moderation design indicated by its opaque decisions, users' moderation experiences show concerns about moderation justice or fairness. Researchers have largely shared a similar argument: improving moderation transparency allows more users to perceive moderation as fairer. Jhaver et al. [44] found Reddit users receiving moderation explanations tended to perceive moderation as fair. Vaccaro et al. [108] also discovered in their experiment that after Facebook users received explanations either generated by humans or algorithms, their perceived unfairness would be relieved. These findings resonated well with the endeavors of designing fairer moderation designs. Both Fan et al. and Vaccaro et al. have tried to involve the ethical value, fairness, in moderation decision-making processes [29] and in contesting moderation decisions [109]. This line of work shows the emerging interests in shaping moderation designs to more consider users' voice and interests.

However, we have not yet fully understood how such emerging interests share findings or visions of moderation research. For example, we observed that communication researchers pay more attention to stressing users' agency given their moderation experiences (e.g., [21,85]), while some HCI researchers focus more on designing just, transparent, and contestable moderation mechanisms (e.g., [29,48,109,112]). And other HCI and computational researchers have aimed to understand whether moderation is effective in suppressing harmful user behaviors (e.g., [16,43,46]). So, it still remains unclear whether and how prior researchers share findings or academic perspectives of moderation experiences. We aim to address these research gaps in this SLR.

2.2 A Systematic Literature Review (SLR) of Moderation Experiences

Conducting a systematic literature review (SLR) is an important approach to systematically understand a research area and identify its gaps and trend, as many previous HCI and CSCW researchers did (e.g., [5,26,98]). Systematically understanding how prior work shares conversations about users' moderation experiences is important. That is because discovering similarities and differences in existing moderation literature could inform a future research agenda for HCI and CSCW as multi-disciplinary fields. Researchers measured how moderation is conducted [45,92], uncovered how content policies were articulated [17,31], designed theoretical models of effective moderation or governance structures [29,88], and developed more transparent moderation tools [48,112]. To enrich this trajectory of moderation research in HCI and CSCW, we see the potential

of conducting an SLR of prior work from HCI, communication, data science, and suchlike through human-centered perspectives.

However, there is nearly no SLR that does so by touching or focusing on moderated users, with one recent exception. In early 2023, Jiang et al. published an SLR focusing on moderation literature published before October 2020 and how these articles depicted the tensions and trade-offs of how moderation is conducted. These tensions include transparency of moderation versing opacity and human moderation versing automated one, etc. [52]. Besides focusing on moderation at operational levels, we still lack and need more knowledge of how prior work depicts users' moderation experiences which are shaped and impacted by these moderation operations. That is because we cannot understand the full picture of moderation design unless we study how users experience it. In other words, we recognize the need to further center moderated users to reflect on moderation design based on moderation experiences. Thus, in this paper, we seek to conduct an SLR to benefit HCI and CSCW communities by suggesting future research agendas for moderation research.

3 METHODS

We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [71] to analyze prior moderation literature to answer our research questions. PRISMA, first published in 2009 by Moher et al. [71], is a standardized guideline including four stages, including identification, screening, and eligibility and inclusion, for conducting and reporting systematic reviews and meta-analyses. It is a widely adopted SLR method that has been conducted in much work of CSCW and HCI (e.g., [5,42,78]) and can also ensure the quality and transparency of systematic reviews, over other scoping methods such as Cochrane Database of Systematic Reviews (SR) [32,70]. We first defined the inclusion criteria of this SLR and then conducted it by following PRISMA in four stages, including (1) *identification,* which helped us search moderation literature in academic databases by relevant keywords, (2) *screening* to screen studies' titles, keywords, and abstracts of identified studies and thus informed our forward and backward snowball sampling, (3) *eligibility* – full-text review on screened results to exclude studies that are not matched with our inclusion criteria, and (4) *inclusion*, which was data extraction and analysis on studies that meet our criteria.

Before investigating all three research questions through PRISMA guidelines, we broke down the second RQ, "How prior work report findings of users' moderation experiences," into a few operational parts. Inspired by the PICOC criteria[2] from existing methodological guidelines of conducting a systematic literature review (e.g., [56]), we distill five WH-questions from the second RQ to cover five empirical aspects of users' moderation experiences, including "who," "what," "where," "how," and "what's next":

RQ2.1: Who are the moderated users described in prior work?
RQ2.2: What moderation decisions do moderated users receive?
RQ2.3: Where do moderated users experience moderation?
RQ2.4: What moderation mechanisms do prior work uncover?
RQ2.5: How do moderated users react to the moderation they experience?

[2] PICOC stands for Population (i.e., a certain population group relevant to technology such as engineer, tester, or users), Intervention (i.e., a technology that tackles a certain issue), Comparison (i.e., the comparison between technologies), Outcomes (i.e., relevant results from technologies), and Contexts (i.e., where the technologies are implemented and compared with each other)

3.1 Inclusion Criteria

The criteria for including studies for this SLR are full empirical studies that investigate how users experience online moderation. We detailed the inclusion criteria in below five points:

- An article must be a full empirical article that is based on observed and measured phenomena to generate insights from actual experiences. The criteria for identifying whether any work is empirical research include (1) answerable research questions, (2) definitive population or phenomena for research, and (3) descriptive processes (e.g., qualitative or quantitative methods) to investigate research questions. Meanwhile, we chose a full paper instead of an extended abstract, poster, work in progress, etc., for our SLR as a quality assessment and control, as much SLR did (e.g., [52,98]).
- An article must be written in English due to the research team's language capacity and as this choice has been performed in the CSCW community (e.g., [5,26]).
- An article must concern online moderation. Online moderation means platform governance mechanisms that regulate the abuses and facilitate community cooperation. We drew from Grimmelmann's definition of moderation because it has been widely adopted by many HCI and CSCW studies (e.g., [44,51,66]). In this definition, the abuses include four general categories, including (1) congestion of infrastructures due to information overuse, (2) cacophony where people can hardly find what content they want, (3) abuse which refers to "bad" rather than information goods, and (4) manipulation, meaning information is skewed.
- An article must be published between January 01, 2016, and March 31, 2022, the date we started the literature search. We set this around a six-year time frame because (1) we aimed to focus more on contemporary work that concerns moderated users and their experiences, (2) a recent SLR concerning moderation [52] found there was a significant increase in the number of papers published in and after 2016, compared to the previous years, and (3) more researchers have started to connect the metaphor, "platform" with content moderation since the year around 2015 or 2016 (e.g., commercial content moderation [83]).
- An article must describe moderated users' experiences. Moderated users' moderation experience refers to users' lived experiences, such as attitudes, behaviors, thoughts, and suchlike through first-person accounts [69,82] after users, either individually or collectively (e.g., user group or online community as a whole), experience moderation.

3.2 PRISMA Stages

Based on the inclusion criteria, we conducted our SLR by following PRISMA [71] in four stages, Section 3.2.1 to 3.2.4, to search, screen, and analyze relevant prior studies.

3.2.1 Identification: Keyword Searching

We identified a search keyword list related to moderation from relevant prior literature with three steps. First, since moderation research has emerged in various fields, including communication [73,99], law [39,59,65], and computational or interdisciplinary areas [17,47,108], the first author of this study who had done empirical studies about moderation extracted terms or phrases from known moderation literature across the fields by reading articles' Introduction and Methods sections.

Table 1. The Boolean search in ACM Digital Library. Please note that two researchers searched both keyword lists in the ten literature databases.

Title, abstract, or keyword contains:		**Filters**
(1) general moderation terms:	(2) specific moderation terms:	"filter": {Article Type: Research Article, Publication Date: (TO 03/31/2022)}
("online moderation" OR "online governance" OR "content moderation" OR "community moderation" OR "platform governance" OR "content regulation" OR "online regulation" OR "platform regulation" OR "community management" OR "moderation strategy" OR "automated moderation" OR "algorithmic moderation" OR "social media moderation" OR "social media content moderation"	OR "moderation decisions" OR "account suspension" OR "screen content" OR "screen user-generated content" OR "comment removal" OR "post deletion" OR "suspended users" OR "comment deletion" OR "content restriction" OR "post removal" OR "content removal" OR "content flagged" OR "account determination" OR "restrict content" OR "prohibit content" OR "reinstate content" OR "reinstate account" OR "account reinstatement")	

For example, recent CSCW moderation literature mentioned "account suspension" and "content removal" in their study context or methodological design (e.g., [41,44,108]), and we identified and named these as *specific moderation terms* because prior researchers used them to describe certain moderation decisions instead of general descriptions of moderation such as "automated moderation." Second, online materials such as news reports [1,115] or community guidelines of social media platforms [37,120,121] further supplemented the search keyword list.

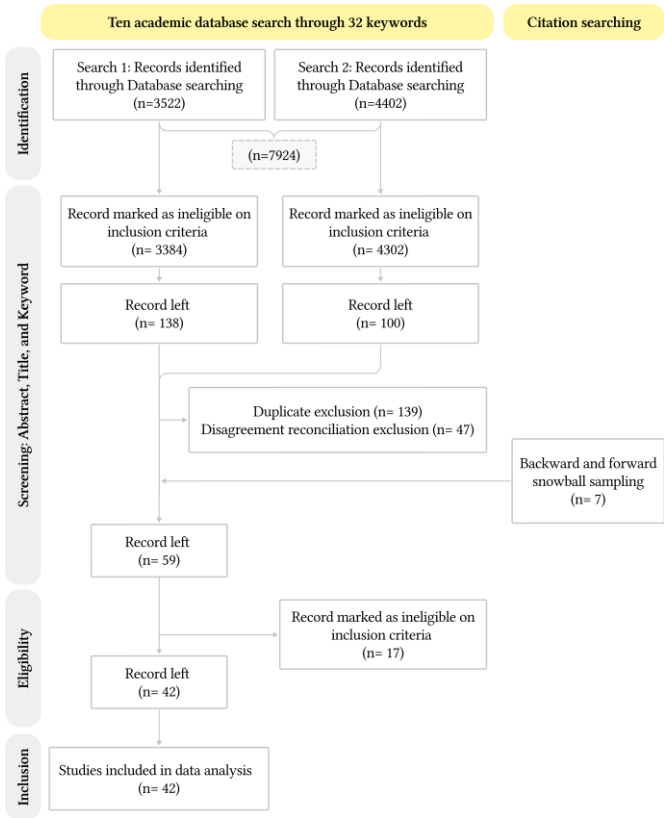

Fig. 1. PRISMA Diagram. Please note that, as we mentioned in Section 3.2.2, we conducted two separate searches to fetch as much literature as possible and then removed the duplicates.

For example, we identified "reinstate account" from YouTube's appeal community guidelines and categorized it into specific terms as it concerns the action that users restore their account after it is suspended. Third, we grouped these two groups of keywords together, general and specific moderation terms. The former refers to terms or phrases that describe moderation in a general way, such as its synonyms, i.e., content regulation or community moderation, and how moderation is conducted (e.g., algorithmic moderation). The latter refers to specific moderation decisions or punishments that end-users experience, such as account ban, content removal, and more (see Table 1).

Ten literature databases were selected to search for relevant moderation literature to ensure our search could cover enough disciplines. The database list included top academic databases that have been widely recognized, such as Scopus, Web of Science, Springer-link, ScienceDirect [122], and computational ones, such as IEEE Xplore Digital Library and ACM Digital Library. As disciplines such as communication and law have actively investigated online moderation, we also included Sagepub, JSTOR, and Taylor & Francis Online for the communication field and heinonline.org for legal scholarship.

Because of such a comprehensive academic database list, we strategically grouped all searching keywords with a Boolean operator, OR, to allow our search to be operatable and search results to be as complete as possible. We detail our search query structure with an example of how we conducted a literature search in ACM Digital Library, as shown in Table 1.

3.2.2 Screening

Given that there might be some function differences in each literature database (e.g., inability to identify quotation marks or Boolean marks) or human errors (e.g., Boolean mark typo), two researchers conducted literature search separately, as Figure 1 showed two searches to ensure that we could gather possibly related literature as comprehensive as possible.

Two researchers then separately examined titles, keywords, and abstracts of in total of 7,924 papers, which included duplicates due to two times of searches, to measure whether they meet this SLR's inclusion criteria. In this process, they settled disagreements between each other by flagging papers that were potentially not relevant to the criteria and discussing them through regular group meetings. Based on this review process, 3,384 and 4,302 papers were filtered out separately, and in total, 238 papers were left. Then two researchers resolved disagreement (N=47) on the initially screened results and removed the duplicates in them (N=139). This step ended PRISMA's *screening* stage by eventually identifying 52 papers.

Forward and backward snowball sampling. Before entering the *eligibility* stage, i.e., full-text review, we conducted both forward and backward snowball sampling to examine how the 52 papers we initially identified were either generated from previous relevant literature or already contributed to further studies. Following the guidelines of iterative citation search articulated by Wohlin [111], we detailed the backward and forward citation search:

- We conducted a backward citation search to identify whether a relevant study was cited by any of these 52 papers. We examined these papers' citation/reference sections to see if any titles, publication venues, or authors could be related to our inclusion criteria. We found one potentially relevant paper.
- For backward citation search, we utilized the 'cited by XXX' in Google Scholar. We identified six papers that potentially fit our selection criteria. Thus, in total, 59 papers entered the PRISMA stage *eligibility*.

3.2.3 Eligibility: Exclusion

The first author screened the full text of each study (N=59) to identify whether they matched the inclusion criteria of this SLR. Seventeen studies were relevant and important to inform future moderation research endeavors, but they did not match our inclusion criteria. We excluded those studies for reasons including (1) studies that have not focused on moderated users' experiences, meaning that some were either not focused on moderation experiences (e.g., behaviors, feelings) (N=7) or not on moderated users (e.g., general social media users, moderators) (N=4); (2) other excluded studies were about how moderation is conducted by algorithms or moderators (N=3) or (3) not empirical ones (N=2) or, (4) not about online moderation at all (N=1). Thus, 42 studies that fit our inclusion criteria were left to enter the last stage of PRISMA, as shown in Figure 1.

3.2.4 Inclusion: Data Extraction and Analysis

Before conducting data analysis, the research team first extracted essential excerpts from each study. These excerpts included author names, publication venues, article titles, and publishment year. Then, they extracted the information regarding (1) moderated users' identities or characteristics that prior studies focused on, (2) what online platforms and the platform affordance described by researchers (i.e., text, video, audio, or image), (3) moderation techniques or modes that platforms used, and (4) the studies' research method uses. This process generated a spreadsheet of extracted data for descriptive analysis to answer RQ1 and partial RQ2.

Two researchers then conducted an inductive thematic analysis [9] on the full text of the 42 studies to answer RQ2 and 3. The researchers read through and familiarized themselves with all literature. Then, they assigned initial codes to literature in terms of how each initial code could represent the portion of literature (e.g., sections, paragraphs, sentences) and potentially can answer research questions in this SLR. Then, they grouped similar initial codes together to form a theme with its definition to answer RQ2 and RQ3.

4 FINDINGS

This section presents how prior literature has investigated users' moderation experiences. It will describe prior work's characteristics (Section 4.1), how users experience moderation and moderated users' reaction stages in the post-moderation phase (Section 4.3), and lastly, how prior studies conceptualize users' moderation experiences (Section 4.3).

4.1 RQ1: Study Characteristics

Prior research (N=42) investigating moderation experiences has grown in a positive trend in the past few years (see Figure 2). Of these 42 studies, exactly half (N=21) were published after the year 2021. Between January 2016 and March 2022, most of the studies were published in the year 2021 (N=18), while only one study was published in February 2016 (N=1), which was the earliest time when we found that contemporary researchers focused on moderated users.

Methodology-wise, slightly over half of the studies (N=23) used qualitative methods such as thematic analysis, content analysis, or digital ethnography on data like interviews (e.g., [12,21,117]), qualitative surveys (e.g.,[73,99]), or online discussions (e.g., [60,66]). The second largest part of prior studies (N=16) used quantitative methods such as interrupted time series regression, clustering algorithms, or running Perspective API to analyze the toxicity of user speech. The last part was mixed-method studies (N=3). These studies all started with user surveys involving quantitative analysis such as binomial logistic regression [41], parametric tests like one-way ANOVA [108], or multiple linear regression [44] to investigate moderated users' behaviors or

perceptions. These three studies all ran qualitative analyses on surveys' open-end questions to uncover users' reasoning or sense-making around moderation decisions.

These studies come from a wide range of disciplines, such as communication, HCI, data science, law, and sociology. HCI and communication have been the two most active fields on this topic, as shown in Figure 3. We identified the field of a study by conditions, including (1) the venue in which it was published and then (2) the academic background of leading authors if we cannot decide by the first condition. Specifically, in the HCI field (N=20), 16 studies were published at the CSCW conference ranging from the year 2016 to 2021, three other studies have been from TOCHI, and one has been made open access on arXiv (at the time when we searched). Among these 20 studies, more than half (N=16) were published after the year 2019, again showing a growing interest in understanding moderated users.

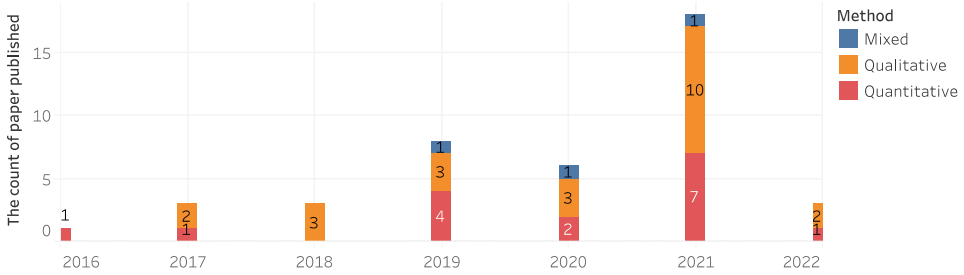

Fig. 2. The growing trend of moderation research (between January 2016 and March 2022) with methods.

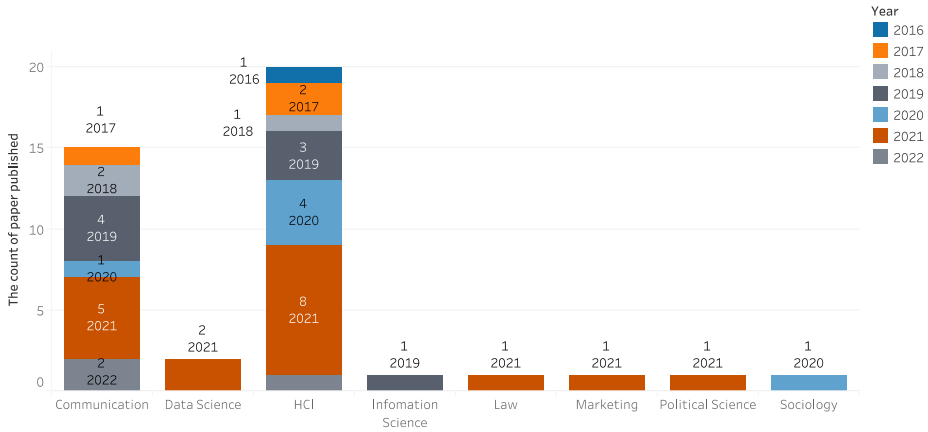

Fig. 3. The distribution of moderation research by research field and year.

4.2 RQ2: Experiencing Moderation

In this section, we discuss in prior moderation literature, (1) who the moderated users were (Who), (2) what types of moderation decisions they experienced (What), (3) on which platforms users experienced moderation (Where), (4) what types of moderation implementation led to users' moderation experiences (How), and ultimately (5) how users reacted to the moderation they experienced (What's next).

4.2.1 Moderated Users

Among the 42 studies we reviewed, one-third (N=14) studied general users, where researchers did not specify user identities but sought to derive general insights into user-moderation interaction.

The rest paid attention to how unique user identities or groups might intersect with moderation actions and yield distinct moderation experiences. These included minority groups (N=11), content creators (N=7), and users whose communities experienced moderation (N=10).

First, minority groups refer to users whose practices, race, or religions are fewer in numbers or more historically underrepresented than other users. Researchers investigated moderation experiences of pro-eating disorder (Pro-ED) users [14,30,33], global south people [24], human-right activists [4], women [74,75] (if we consider that they have less power than men [40]), early adolescents (i.e., ages 10 to 13) [101], and sexual or racial minority groups [34,41,109]. These researchers have been concerned with pro-ed users' struggles in self-recovery due to moderation and their lack of social support [30], disproportionate account suspension and content removal that happen to sexual or racial minority people [41,109], and the prioritized appeals initiated by some users than others [4].

Second, researchers focused on creative labor like content creators (N=7). They are typically influencers on Instagram [21], YouTubers [10,12,55,66], and creators on TikTok [85,117], contributing to the platform economy.

Lastly, a relatively special group of moderated users is those affected by moderation decisions taking place on their online community as a whole (N=10). Prior work investigated how they experience community-level (N=9), such as subreddit restriction or ban, as well as platform-level moderation (N=1). Community-level moderation means that platforms moderate specific online communities. Nine work uncovered how users might behaviorally respond to such moderation. Platform-level moderation refers to platforms' content policy changes or moderation practices that apply to users, including all user communities, universally. One work focused on how users experience and push against platform-level moderation globally [100].

4.2.2 Moderation Decisions that Users Experience

Most studies investigated moderation experiences with speech or text-level moderation (N=19), meaning that user's content (e.g., a post created on Facebook) is directly removed (N=15) [4,14,18,24,30,33,34,41,44,47,61,73–75,91,95,99,106,109] or prevented to be published (N=4), e.g., hashtag ban [14,33], in-streaming or in-game chat restriction [61,91]. Of the 15 studies that investigated experiences with content removal, six [4,41,61,73,99,109] included discussions of how users experience account suspension. Thus, Figure 4 shows the number of total literature greater than 42. Besides, two studies focused exclusively on account suspension [60,108].

Besides fifteen studies focusing on text or speech moderation, growing research (N=7) has investigated experiences with the restrictions on the status of user accounts while the content remains. Such moderation includes algorithmic restriction (e.g., content/account visibility decrease), economic restriction (e.g., monetization capability deduction), community engagement restriction (e.g., liking/commenting decrease), and more. For example, researchers (N=3) [21,85,117] explored how platforms algorithmically restricted the visibility of content and prevented audiences from finding creators' content, the moderation some researchers framed as "shadowban." Others (N=4) [10,12,55,66] investigated "demonetization," where content creators encountered a decrease of the future income generated by content.

Furthermore, a few studies (N=4) focused on platform-specific moderation decisions, which we called "other moderation" in Figure 4. Specifically, the moderation decisions such as block-listing users on Twitter [49], removing users from online communities on Minecraft [101], issuing warnings to users on Twitter [114], and removing users' moderator positions in online communities [113] shaped various users' perceptions or behaviors.

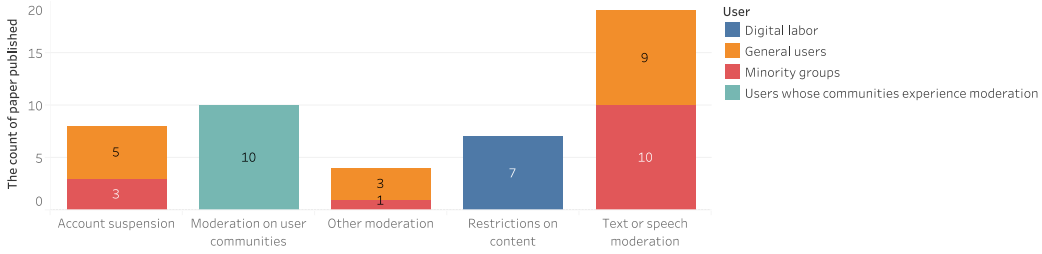

Fig. 4. Moderated users and the moderation decisions they experienced.

As explained in section 4.2.1, ten studies [15,16,19,43,46,81,100,102–104] focused on users' experiences with the moderation that took place on their communities. Nine studies investigated experiences with subreddit ban or quarantine [15,16,19,43,102–104], Facebook page removal [81], influencers' permanent account suspension [46], and moderation practices change on Tumblr universally [100].

4.2.3 Platforms that Moderate Users

An important research question of this SLR is where users experience moderation. As shown in Table 2, we found most researchers investigated moderated users on platforms primarily affording (1) textual or speech content (N=25), (2) both texts and images (N=6), or (3) texts, images, and videos (N=2). Among these three focuses of platform in prior work, most focused on specific platforms such as Reddit (N=12), Facebook (N=5), Twitter (N=3), YouTube (N=3), and Instagram (N=3).

4.2.4 Moderation Implementation Matters to Moderation Experiences

Platforms today rely more on either centralized, commercial, or community-based, voluntary initiatives to moderate users. Platforms such as Facebook, Twitter, Instagram, YouTube, or TikTok, enforce centralized and platform-wide rules to "guard against digital damage to their brand" from problematic content through commercial services of human workers [83] or algorithms [38]. Platforms value moderation as business decision-making for brand images or clients' stake. For example, YouTube removes and demonetizes (i.e., deducting future advertising income) hundreds of YouTubers because inappropriate viewers' comments under those YouTube channels harm the brand images of both advertisers and platforms [12,96].

While many platforms conduct such commercial moderation, some platforms comprised of various communities (e.g., subreddits) implement community-based, voluntary moderation through community members. Each community might have contextualized moderation designs such as community rules [17,31], moderator structures (e.g., [45]), and thus different definitions of how acceptable a user behavior is. This means community-based moderation is not always consistent across platforms. For example, the same users might experience less force of moderation and thus receive more freedom of speech in one subreddit than the other subreddit, which has stricter content rules [34].

These distinctions between commercial and community-based moderation imply the need to understand users' moderation experiences under different moderation initiatives. For instance, in community-based moderation, human moderators in communities such as subreddits or Twitch channels can remove users' content or accounts, as explained in Section 4.2.1. In commercial moderation that is applied to platform-wide users universally, TikTok can deduct creators' visibility across countries such as the United States, Canada, and Australia [117]; Tumblr changed

platform-wide content policies to every online community (e.g., discussions around hashtags) on it [100], and YouTube can remove YouTubers' videos and also decrease or remove their monetization capabilities globally [12,55,66]. Based on the differences between these two types of moderation initiatives, we identified that more studies focused on how users experience commercial moderation (N=33) than community-based one (N=9), as shown in Table 2.

Furthermore, in prior work, users primarily experienced an ex-post reactive mode of moderation (N=39). Ex-post reactive moderation [39] means that user behavior or content is reviewed after it is flagged by either algorithms or human workers [57]. And platforms nowadays adopt such mode widely [93]. Less frequently, some platforms (e.g., Twitch) could alternatively set keyword detection to alert or prevent users from posting something violating community rules [91]. This type of moderation, i.e., moderating content before it is posted, is recognized as ex-ante moderation [57] and appeared in three prior studies.

Given this research background, we recognized the need to understand how users' moderation experiences reflect what techniques the platforms operate in moderation designs, as we summarized in the column "technique" of Table 2. Most prior studies (N=18) explicitly indicated the combination of human and automated tools (e.g., algorithms). Human moderation, either voluntary community members or commercial moderators, takes a role in flagging, reviewing, and moderating a user who is deemed to violate platform rules [23,36,39]. Also, platforms usually implement algorithms in moderation [2,23,45] along with human moderators' support. Based on this background of humans working with automated moderation, we found (1) seven studies explicitly focused on automated moderation, (2) seven studies uncovered users' experiences with human moderators, and (3) ten studies did not explicitly describe whether users experience automated or human moderation.

4.2.5 Reaction Stages after Encountering a Moderation Decision

By analyzing 42 studies, we found that prior work commonly presented how users react to moderation decisions in three primary stages, including users' emotional responses, cognitive processes where users generate perspectives or opinions about moderation and behaviors informed by perceptions or emotions, the prior two stages, as shown in Figure 5.

Emotions. Prior work described users' different emotions regarding moderation decisions. While human emotion is a subjective, internal experience, we leveraged Plutchik's categorization of eight basic emotions (i.e., joy, trust, fear, surprise, sadness, anticipation, anger, and disgust) as keywords to effectively identify and analyze how prior work describes moderated users' emotions, especially those negative ones [79,80], to moderation decisions.

We found five major emotions that prior studies commonly reported. The most salient emotion, *frustration or confusion*, appeared in eight studies [4,10,30,44,66,73,85,117]. Frustration refers to the feeling of upset because users felt unable to change the fact that they experienced moderation. Prior work described that this emotion frequently appeared with users' confusion to further depict how users felt it hard to make sense of such fact happened to them. For example, users on Reddit felt more frustrated at the lack of notification of content removal than the removal itself [44].

Table 2. Prior studies describe what platforms and how they implement moderation.

Moderation experience	Platform		Moderation		
Studies	**Platform Focus**	**Moderation Focus**	**Technique**	**Mode**	**Initiative**
Banchik [4]	Multiple platforms	Text & Image & videos	Human and automated	Ex-post	Commercial
Brøvig-Hanssen & Jones [10]	YouTube	Video	Automated moderation	Ex-post	Commercial
Caplan & Gillespie [12]	YouTube	Video	Automated moderation	Ex-post	Commercial
Chancellor et al. [14]	Instagram	Text	N/A	Ex-post	Commercial
Chandrasekharan et al. [15]	Reddit	Text	Quant	Ex-post	Commercial
Chandrasekharan et al. [16]	Reddit	Text	N/A	Ex-post	Commercial
Christodoulides et al. [18]	N/A	Text	Human	Ex-ante	Commercial
Copland [19]	Reddit	Text	N/A	Ex-post	Commercial
Cotter [21]	Instagram	Text & Image	Automated moderation	Ex-post	Commercial
Das et al. [24]	Quora	Text	Human	Ex-post	Community
Feuston et al. [30]	Multiple platforms	Text	Human and automated	Ex-post	Commercial
Gerrard [33]	Instagram	Text & Image & videos	Human and automated	Ex-post	Commercial
Gibson [34]	Reddit	Text	Human	Ex-post	Community
Haimson et al. [41]	Multiple platforms	Text & Image	Human and automated	Ex-post	Commercial
Jhaver et al. [44]	Reddit	Text	Human and automated	Ex-post	Community
Jhaver et al. [46]	Twitter	Text	N/A	Ex-ante	Commercial
Jhaver et al. [47]	Reddit	Text	Human and automated	Ex-post	Community
Jhaver et al. [49]	Twitter	Text	Human and automated	Ex-post	Community
Kaye & Gray [55]	YouTube	Video	Human and automated	Ex-post	Commercial
Kou [60]	League of Legends	Online game	Human and automated	Ex-post	Commercial
Kou & Gui [61]	League of Legends	Online game	Human and automated	Ex-post	Commercial
Ma & Kou [66]	YouTube	Video	Automated moderation	Ex-post	Commercial
Nurik [75]	Facebook	Text	Human	Ex-post	Commercial
Procházka [81]	Facebook	Text	Human and automated	Ex-post	Commercial
Ribeiro et al. [43]	Reddit	Text	N/A	Ex-post	Commercial
Savolainen [85]	TikTok	Video	Human and automated	Ex-post	Commercial
Seering et al. [91]	Twitch	Text	Automated moderation	Ex-post & Ex-ante	Community
Srinivasan et al. [95]	Reddit	Text	Human	Ex-post	Community
Suzor et al. [99]	Multiple platforms	Text & Image	Human and automated	Ex-post	Commercial
Sybert [100]	Tumblr	Text & Image	Automated moderation	Ex-post	Commercial
Tekinbaş et al. [101]	Minecraft	Online game	Human	Ex-post	Community
Thomas et al. [102]	Reddit	Text	N/A	Ex-post	Commercial
Trujillo & Cresci [103]	Reddit	Text	N/A	Ex-post	Commercial
Trujillo et al. [104]	Reddit	Text	N/A	Ex-post	Commercial
Tyler et al. [106]	Facebook	Text & Image	Human and automated	Ex-post	Commercial
Vaccaro et al. [108]	Facebook	Text	Human and automated	Ex-post	Commercial
Vaccaro et al. [109]	Participatory design	Text & Image	Human and automated	Ex-post	Commercial
West [73]	Multiple platforms	Text	Human and automated	Ex-post	Commercial
West et al. [74]	Facebook	Text	Human and automated	Ex-post	Commercial
Yang [113]	Reddit	Text	Human	Ex-post	Community
Yildirim et al. [114]	Twitter	Text	N/A	Ex-post	Commercial
Zeng & Kaye [117]	TikTok	Video	Automated moderation	Ex-post	Commercial

Moderation researchers uncovered users' *sadness* (N=5) [10,24,30,44,49] and even a feeling of "unworthy to be seen," as described by Feuston et al.' study focusing on Pro-ED users [30]. The *sadness* also appeared after Bangladesh users interacted with moderation teams of Quora, where moderation practices tended to privilege the dominant national and religious identities [24].

Besides the users' *sadness*, four prior studies [4,30,44,108] presented users' *anger*. *Anger* refers to a strong feeling of annoyance due to perceived unfair moderation. For instance, users became angry when they tried to initiate their appeals for account suspension decisions that they perceived as unfair [108].

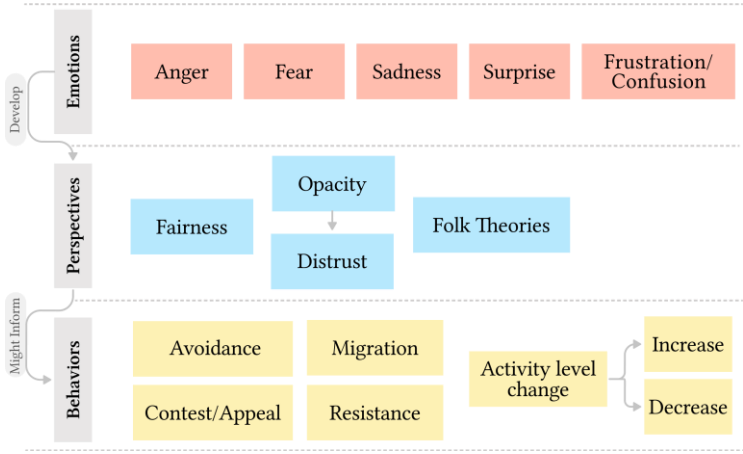

Fig. 5. Users' three-stage reactions model after they experience moderation.

Moreover, *surprise* (N=2) refers to a mismatch between users' understanding of moderation and how they actually experience moderation. For example, users felt it surprising that their posts were removed due to a clause of content rules that violated their common sense [44].

Lastly, *fear* (N=2) means the emotion of being afraid to experience pain or a threat. For instance, YouTubers feared that appealing a video takedown would expose them to greater risk than simply accepting it [10].

Post-moderation Perspectives. If negative emotions are users' immediate, instinctive responses to a moderation decision, then some users also move beyond this emotional reaction stage, conduct informational work around the moderation decision and the moderation design, and formulate informed understandings and perspectives. For example, YouTubers initially felt frustrated about moderation and its impacts on decreasing their ad income [66] and viewer engagement [10]. And then, they understood moderation algorithms were opaque due to their lack of knowledge of how moderation is enforced. Such processes that emotions develop to users' perspectives appeared in a bunch of work, and we identified four types of post-moderation perspectives.

Fairness perceptions (N=8) and perceived opacity (N=8) of moderation, separately, were the two conspicuous perceptions that prior researchers uncovered. Fairness perception means whether and how users feel moderation is fair [12,30,41,44,49,75,109]. For example, users felt it was unfair that they had not been active on Twitter for a long time while suddenly finding out they were on a blocklist [49]. Also, some women experiencing content removal on Facebook felt it unfair that they could not post anything mean about a man, even jokes [75].

Besides, eight studies [4,24,30,60,61,66,73,108] uncovered the perceived opacity of moderation, especially its decision-making. Users from different online or social identities generate such perceptions. For example, players in the game League of Legends requested explanations that could help them understand what their past behaviors were identified as offensive [60], and that can further help reform their problematic behaviors [61]. Either users who experienced account suspension on Facebook [73,108] or content creators who experienced demonetization [66] requested detailed reasons, such as the reference to content policies.

This perceived opacity of moderation could be further connected with users' distrust of platforms (N=4) [44,60,99,108]. For example, Suzor found that moderation happening without notifications induced users' distrust [99]. Kou also uncovered that players held distrust of the

game publisher when it issued inconsistent punishments without sufficient explanations [60]. As Jhaver et al. stressed, moderation explanations should be a helpful and instructional design for platforms to gain moderated users' trust [44].

Lastly, we found that moderated users generated different folk theories about moderation in prior work (N=5). They believed that moderation was politically biased against certain people's voice [44,85], understood that specific keywords would trigger moderation algorithms [66], or thought platforms selectively recommended or moderated some users' content [55,117].

Post-moderation Behaviors. Prior moderation research largely uncovered users' behaviors after they experience moderation, and we call such behaviors post-moderation behaviors. We identified primarily five such behaviors, as shown in Figure 5.

Unsurprisingly, prior work (N=10) [10,14,30,33,34,55,61,81,101,117] found many users tended to avoid or circumvent moderation. Users might alter (1) linguistic characteristics or clips of their user content [14,30,33,34,55,81,101], (2) post content that could lower the possibilities of getting attention by platform algorithms (e.g., searching) [10], or (3) turn off in-game chat function to avoid becoming toxic in online games [61]. Similar to avoidance actions, prior studies (N=6) found that users either switched between the original platforms they experienced moderation and alternatives [66,113] or left the original ones [16,19,102,108].

In contrast to migration and avoidance, many studies (N=11) discovered that users resisted or pushed back moderation [4,30,106] (e.g., posting complaints about platforms' content policies [100]). Related to resistance, what many moderated users did is to initiate appeals to request platforms to re-review moderation decisions (N=10) [4,10,30,49,61,66,73,106,108,117]. However, prior studies also found that users perceived appeal processes as opaque [30,66,108,117] or even unfair [49]. So, moderated users tended to contact their platform administrators (e.g., human moderators) through third-party platforms [66,73] or personal connections [4] to better contest moderation decisions.

Another key theme of post-moderation behaviors in prior work (N=21) is the noticeable changes in user activity or engagement level. In detail, users might be active in posting content or writing hate speech online before they experience moderation. However, the extent of such activities might present in either increasing or decreasing trends after users experience moderation. For example, seven studies [14,19,43,47,95,103,117] described how users, after moderation, became more active in generating content [47,95] (e.g., pro-ED posts [14]), attached content with more hashtags [117], or even became more toxic [103] and hostile [43]. Compared to behavioral level increase, 14 studies [15,16,19,43,46,47,61,91,95,103,104,106,113,114] uncovered users' activity level decrease. For instance, after community-wide moderation (e.g., subreddit quarantine or shut-down) happened, fewer newcomers joined such communities, and veteran users became less active and toxic [15,16,19,43,46,104]. Generally, after experiencing moderation, users were found to decrease their actual content-generating frequency or tendency [47,95,106,113] and also their toxicity, including spamming [91], hate speech [16,114], and suchlike [61].

4.3 RQ3: Conceptualizing Users' Moderation Experiences

In this section, we will discuss five general perspectives that prior research drew upon to frame or conceptualize moderation experiences: 1) The effect perspective that emphasizes measuring the effect or effectiveness of moderation decisions; 2) The agency perspective that concerns user agency or how users, individually or collectively, exert agency to deal with moderation decisions through actions such as appeal, resistance, organizing, and sense-making; 3) The ethical

perspective that draws from ethical values such as fairness and transparency to interpret moderation experiences; 4) The marginalization perspective that examines how moderation techniques systematically marginalize minority groups, and 5) The creative labor perspective that primarily focuses on content creators' moderation experiences.

4.3.1 The Effect Perspective

What falls into the effect perspective is a group of studies (e.g., [14,15,19,46,91,95,102,104]) dedicated to measuring the effects of moderation decisions. These studies tend to be guided by questions that explicitly state researchers' interests in moderation effect. For example, Chandrasekharan et al. asked, "what effect did Reddit's ban have on the contributors to banned subreddits?" in their study of analyzing the effects of subreddit ban on users' hate speech [16]. For another example, Seering et al. tested their hypothesis on whether chat moderation modes on Twitch could decrease the frequency of spam and found that fewer spam messages appeared after such moderation [91]. This group of studies generally tends to use quantitative metrics to measure moderation effects. For instance, Jhaver et al. used causal inference methods to test whether anti-social ideas and discussions around influencers would decrease after these influencers were deplatformed by Twitter [46].

The effect perspective aligns with the view of moderation as a problem-solving scenario, where the effect of the solution naturally occupies the central place in evaluating moderation. Only a few of these studies have also paid attention to users' subjective experiences. Jhaver et al. studied users' perceptions of content removal they experienced on Reddit. The authors found that users felt frustrated at their content removal and complained it was unfair to violate their freedom of speech [44]. This lack of mixed-method work focusing on moderation effects indicates more need for not only quantitatively measuring moderation effects but also triangulating quantitative findings with how users perceive or react to moderation.

4.3.2 The Agency Perspective

If the effect perspective treats moderated users as a subject acting in accordance with moderation decisions, the agency perspective pays attention to users' capacity to act on their own when affected by moderation decisions. For instance, communication and other social science researchers (N=12) have stressed how moderated users can act against existing moderation practices and content rules. Sybert found that although users are in lack of user agency and free expression under the over-policing of their bodies and sexuality, they combated and undermined the platform owners' authority by generating memes to critique, resist, and satirize [100]. Beyond individual actions, West et al. found that moderated users successfully leveraged both newcomer protesters and celebrities to obtain social media platforms' attention to refining content rules. And the authors have theorized that moderated users can strategically leverage their collective agency to combat the platforms' power [74].

4.3.3 The Ethical Perspective

In line with rising ethical concerns regarding contemporary algorithmic systems, especially those enhanced by artificial intelligence techniques, moderation researchers value ethical values in moderation systems and use them as an interpretive lens to understand moderation experiences. For example, Vaccaro et al. have focused on soliciting different ethical values of moderation designs, including transparency, fairness, and accountability, with moderated users [108,109]. Jhaver et al. have also distilled the transparency of community-based moderation on Reddit from how users encounter moderation explanations [44,47]. Ma and Kou have further stressed the

fairness that moderation systems should be manifested in moderating content and deducting YouTubers' advertising income [66].

4.3.4 The Marginalization Perspective

Our findings, as described in Section 4.2.1, suggest that researchers (N=8) have increasingly paid attention to marginalized groups in terms of how moderation affects them. These researchers tended to unearth how (1) gender or sexual minority groups experienced, perceived, made sense of, or handled disproportionate moderation decisions, (2) women's expressive potentiality was suppressed, (3) Pro-ED community members conformed or circumvented moderation, and (4) users with minority linguistic identities or practices were pushed to the margins.

This list of topics indicates the collective argument that online platforms should be equitable and inclusive for users. Nurik has argued that social media platforms like Facebook prioritize certain user profiles over protecting users' freedom of expression, and women who have been historically marginalized are negatively impacted [75]. In more recent HCI research, Vaccaro et al. have supported that racial and sexual monitory groups should co-construct moderation systems with platforms to improve these users' representation and inclusion [109].

4.3.5 The Creative Labor Perspective

Increasingly, researchers (N=7) aim to study how creative labor like content creators who create content to promote their own media brand [22] experience moderation and how moderation affects their work and life (e.g., income, audience community, and livelihoods). Content creators such as YouTubers, creators on TikTok, or influencers on Instagram might encounter moderation that constrains their visibility [3,7,21], identities [7], or revenue [10,12].

Given such restrictions, many researchers have seen creators' moderation experiences as negotiations between creators' self-interests and platforms' business interests. For example, Caplan and Gillespie have theorized that YouTubers signing the YouTube Partner program [116] are not a "partner" to the platform but enter a new form of contract that treats YouTubers unequally through demonetization (i.e., advertising income deduction) [12] Kaye and Gray have iterated that moderation on YouTubers is in a structural bias that favors and grants more power to larger media organizations than general YouTubers with small fanbase [55].

5 DISCUSSION

Growing work has investigated users' moderation experiences with an expanding scope of interests. From this growing body of work, we identified 1) moderated user types, 2) the moderation decisions they encountered, 3) platforms where they experienced moderation, 4) moderation implementation that shaped their moderation experiences, and 5) the three-stage reaction model after users experience moderation. We further found how researchers from different fields, such as communication/media studies and HCI, conceptualized the empirical findings of moderation experiences differently.

Given our findings, this section will discuss how we compare between early-day's community management design and contemporary platform moderation design and further reflect on moderation design, especially its negative consequences on users, given the empirical studies in this SLR. We further describe how to treat moderated users as a relevant stakeholder group in moderation design to inform future research agendas.

5.1 Reflecting on Moderation Design through Empirical Account of User Experience

Moderation has been a topic of interest for many decades. In the 1990s, when Internet users started to socialize and find communities online, moderation research primarily concerned "regulating deviant behaviors" [11,63] or "community management," such as adjudicating "cyber harassment" cases between harassers and victims within an online community context [97]. How to conceptualize the online context to be moderated has gone through changes, as researchers gradually picked up the notion of "platform" after the 2000s [35]. The appearance of a metaphor, "platform," implies how fast-growing, technical platforms like Facebook and YouTube as content-hosting intermediaries extend and advance the meaning and operation of moderation. First, besides community management's primary purpose of maintaining the productiveness and commitment of users [63], platform moderation further concerns its promise of political neutrality, where platforms grant users with "freedom of speech," which is also substantially questioned by many researchers [35,36,57]. Second, given the sheer increased volume of user-generated content after the 2000s, platforms progressively implemented more advanced technologies (e.g., machine learning [38]) or human-machine collaboration [45] in content moderation. This indicates that compared to general sociotechnical design in community management (e.g., manual moderation or content filtering), platform moderation further presents the innovation and scale of technologies as its signified image [35]. Third, deviant behavior on platforms is not as simple in its taxonomy as what community management targets. Community management primarily tackles textual user content [63], while various moderation challenges emerge with video [10,66], images [30,33,100], and audio [51] types of content on platforms, as summarized in Section 4.2.3. Especially, deviant behaviors are diverse at a conceptual level with the four categories we discussed in Section 3.1 and at an operational level with novel types of problematic content such as adult materials and terrorist content [72,100,107]. Such plural taxonomy of deviant user behaviors further implies the extensiveness of platform policies, which not only take users who generate content into account but also other stakeholders such as advertisers [12,58,67] who might be impacted by problematic user content.

However, our SLR shows that such extensiveness and innovation of platform moderation design do not always successfully generate effectiveness for maintaining productive communities. Rather, prior work concerning moderation experiences uncovered and reflected six negative, unintended consequences of moderation on online community members:

First, moderation tends to assume its authority in decision-making, allowing little to no room for negotiation or contestability. For example, Vaccaro et al. found that users contested Facebook's inconsistent and opaque moderation practices because users found moderation decisions and explanations difficult to understand and interpret [108]. Such deficiency of moderation design echoed well with the design claims that many prior researchers such as Kraut & Resnick and other colleagues [63] stressed, including consistent moderation criteria/standards, more chances to appeal moderation decisions, and moderation decision-making conducted by online communities with rotating power. If online platforms took these prior design claims into account in designing moderation algorithms, moderated users would not ever encounter algorithmic moderation decisions conflicted with content rules [55,68] or lengthy procedures of appealing the decisions [67,73,109], as recent researchers uncovered.

Second, moderation tends to structure platforms and their users as opposing parties. In such logic, platforms typically use ex-post reactive moderation modes to flag and punish identified "bad" users while scarcely hearing from users' voice on what they want to contribute to moderation designs and what they wish to receive to become sound community members. To

achieve this, there is still room for exploring alternative designs of moderation, given moderation experiences. It might be moderated user education instead of punishment each time [73], encouraging desirable user behaviors [91], or testing how users react to and perceive different moderation designs, as we identified in Section 4.2.4.

Third, moderated users are left on their own and under-supported. Prior literature clearly shows how users develop negative emotions (e.g.,[4,10,30,44,66,117]), have trouble making sense of their penalties (e.g., [66,73,108]), and struggle to reform behaviors (e.g., [60,61,66]). These ramifications of moderation align well with the different design advocations researchers have made consistently. For example, Jhaver et al. have called for moderation explanations to be grounded with content rules for better moderation transparency [44]. Vaccaro et al. have further solicited different ethical values of moderation systems with historically marginalized users to ideate better moderation design [109].

Fourth, moderation induces some far-reaching effects on users. As shown in Section 4.2.5, users who experienced community-level moderation might generate animosity toward other users or transfer to less restrictive platforms [19] and become more active (e.g., more posts) [43]. Thus, community-level moderation might not always effectively regulate and educate individual users on their original platforms. Rather, it shapes some users to become more toxic and post polarized content [103].

Fifth, moderation can perpetuate the existing social inequality. As we identified in Section 4.3.4, women and racial, gender, and sexual minorities experience disproportionately more moderation than others [41], and platforms can privilege people with certain races or clans over others [24].

Last, moderation might create extra burdens for users beyond disciplining them. For example, content creators need to go through to piece together their moderation experiences to make sense of why moderation happens, collaborate with other creators to learn to avoid moderation, and switch to alternative platforms to gain more stable income [10,66]. Even moderation produces tension between users and platforms, and users need to leverage collective action efforts to push platforms to refine content rules [74].

Taken together, prior work has pointed to the social side of moderation as experienced by platform users. Initially implemented as a solution, moderation inevitably has a ripple effect on the user community through ramifications that are not yet designed for. Moving beyond this solutionist paradigm, HCI and CSCW researchers have attempted or proposed alternative modes of moderation from users' moderation experiences, such as restorative justice approaches or representative moderation, which empowers users to influence content rule articulation and moderation decision-making [109]. Still, there could be more efforts to understand what other moderation designs work better for users.

5.2 Designing Moderation Experience with Users

Prior work helps carve out a space for designing moderation experiences. Designing for moderated users means moderated users could be one group of users for sustaining and supporting platform governance. Although prior work portrayed that "bad" users (e.g., harassers on Twitter [49]), growing work has also reflected on the deficiencies of moderation design and discussed how users might have not been moderated in the first place [66,73,99,109], as reported in Section 4.2.5 and 4.3.2. Thus, it is valuable and necessary to hear moderated users' voice and see how it could help reflect on and refine existing moderation design (e.g., [109]).

5.2.1 Connecting Moderation with Users

As we identified in Section 4.2.4 and 4.3.2, moderation might not always perfectly reach its effectiveness in reducing unacceptable behaviors while users might behave unexpectedly. They might become less active in generating content and engaging with community members (e.g., [47,95,106,113]), indicating that moderation might unexpectedly shape them to be less engaged.

Platforms, thus, should sufficiently communicate with moderated users. The restorative justice approach leveraged by prior moderation research has stressed the necessity of communicating with and understanding the needs of users who have been offended or harmed and addressing such harm together with offenders [89]. However, as reported in Section 4.2.2, the so-called "offenders" who are deemed to violate platform rules are not always the users who create harm on platforms. These users experiencing "demonetization" [21,85,117] or "shadowban" [10,12,55,66] encounter harm imposed by platforms, oftentimes unexpectedly or false-positively. They need to make through such harm impacting their livelihoods [66], engagement with other users, motivations for generating new content [55], and suchlike due to moderation decisions.

Future work could use alternative justice framework such as restorative justice approach to understand how moderation influences users. Especially prior work reported that users receive moderation due to imperfect moderation mechanisms (e.g., opaque moderation algorithms [21,38,84], human moderators' limited knowledge of cultural contexts of content [99]). Such research efforts could directly help platform owners reflect on existing moderation mechanisms from moderation experiences to prevent unnecessary harm from happening to users.

Prior researchers derived design implications from moderation experiences to suggest more transparent and fair moderation designs (e.g., [41,61,66,108]). However, relatively little attention has been paid to involving moderated users in directly designing effective moderation mechanisms. One exception work done by Vaccaro et al. is that the authors have organized participatory workshops with moderated users from marginalized groups to brainstorm what ethical values moderation systems should contain [109].

Then new questions surface: Do platforms enable moderated users to communicate with moderation designers? Could moderated users collaborate with platform owners/representatives to collectively articulate content rules aligning with their localized contexts? The incentives of these questions are from what we discussed in Section 4.2.5: Moderated users complained moderation decisions failed to be issued with reference to content rules (e.g., [61,73,99]) while platforms offered inconsistent definitions of online harm at the same time [77].

Considering moderated users as relevant stakeholders, future research could explore how they could collaborate with policymakers, designers, or platforms.

5.2.2 Taking into Account Moderation Literacy

The findings of this SLR that researchers started to investigate how users generate folk theories on moderation decisions (e.g., [66]) point to an important research path of understanding users' moderation literacy. We define moderation literacy as users' capabilities to understand and learn about moderation. Users could learn from their own moderation experiences or others to increase such capability, as reported in Section 4.2.5. This shows that users still have a certain extent of agency under authoritative platform governance [21,117].

To support users, especially those who believe to be falsely moderated, future research could study how users experience the learning aspects of moderation design. That means, how users could learn to self-regulate their future behaviors. For example, does user's moderation experience on one platform inform their practices on other platforms?

Furthermore, along with calls on improving moderation transparency to educate moderated users [44,73], relatively little attention has been paid to how moderated users hope the ways they want to be educated. Users could exert their agency to decide whether they want to be productive community members.

5.2.3 Building Mutual Trust between Moderated User and Platform

However, treating moderated users as stakeholders does not mean it applies to every such user. For example, when information about how algorithms work goes public, users might game or misuse algorithms [25], similar to how players appropriate flag mechanisms for competition and achievement purposes in game [62]. It is thus reasonable for platforms to question moderated users' trustworthiness. Similarly, users might not trust back platforms due to moderation (e.g., [60,108]). As such potential mutual distrust grows, we have little knowledge of whether platforms and users need to build mutual trust. If so, have platform owners already defined trusted users, and how could they increase users' trust in platforms? These research questions around mutual trust, especially on its establishment process from the perspectives of platform (e.g., human reviewers or moderators), would be valuable for future moderation research.

5.3 Implications for Future Moderation Research

Efforts are needed to test the effectiveness and utilities of alternative moderation designs. As we summarized in Section 4.2.4, platforms primarily conduct ex-post reactive moderation. For example, much evidence is shown that online communities conducting ex-post reactive moderation can successfully restrict problematic users [45,54]. However, prior work has further pointed out that flagging mechanisms, as an ex-post reactive moderation, can be opaque and be appropriated by problematic users [23] to prioritize self-achievement [62]. Meanwhile, we have not yet fully understood whether ex-ante moderation could help repair the drawbacks of ex-post reactive moderation. While one study conducted by Seering and colleagues found users' problematic behaviors could be relatively successfully restricted in different moderation modes [91], we are not sure how users personally perceive the ex-ante moderation mode compared to the ex-post reactive one. Would users perceive ex-ante moderation as fairer and more transparent? If so, what factors do users generate such perceptions? Would there be differences in moderation experiences by combining ex-ante and ex-post reaction moderation together? The answers to these questions could be valuable to assess whether the existing widely adopted moderation mode, ex-post reactive moderation, is ideal for constructing a more transparent and fair moderation design.

More needed work is on designing fair and contestable moderation for digital or creative labor. For example, content creators, as reported by communication or media study researchers, experience shadowban and income deduction, which frequently takes place on platforms that afford video, audio, and more (e.g., Instagram, YouTube, TikTok) [20,66,117], beyond how users experience content removal within the text contexts. Also, we call for more work on how other digital workers, like human moderators, experience moderation. As this group of users helps conduct and are knowledgeable about moderation, their identity plurality as both moderated users and moderators might help design fairer and more contestable moderation systems, as some CSCW researchers have called for (e.g., [29,44,109]).

More efforts can investigate marginalized people's moderation experience and design inclusive moderation. There are relatively limited deeper insights into how marginalized users perceive and interact with moderation systems. For example, we have little understanding of whether there is a difference in marginalized people's moderation experiences between

video/audio-focused platforms and text-focused ones. Especially when we consider women as culturally less powerful than men [40], HCI researchers have paid relatively little attention to their moderation experiences compared to the only one from the communication field [75].

6 LIMITATIONS

Our systematic literature review is not perfect without shortcomings. First, the search engines of academic databases might have their own limitations. While we extensively searched in ten academic databases, these databases might not always fetch and offer consistent results (e.g., ACM Digital Library [87]). Thus, two researchers fetched the literature data separately to ensure our data collection was as comprehensive as possible and conducted forward and backward snowball sampling to ensure to review moderation research as comprehensively as possible. Second, the studies in this review were limited to papers written in English. As moderation or online harm could be perceived differently by people from different geographic regions [53], future research could seek to review moderation literature written in different languages.

7 CONCLUSION

This SLR focuses on prior literature that investigates how users experience online moderation by synthesizing 42 empirical studies. This review shows that prior studies have built up expansive conversations around moderation experiences, and researchers conceptualize the findings of users' moderation experiences differently. We reflect on platform moderation design, especially its deficiencies, given these findings and further stress that beyond a punitive, solutionist logic, where platforms flag and punish identified "bad" users while scarcely hearing from users' voice, we argue there is still room for future work to explore how alternative moderation design (e.g., platform affordance and moderation implementations) could better shape online communities. We conclude with ways of how to treat moderated users as stakeholders in moderation design and implications for future moderation research.

8 ACKNOWLEDGMENTS

We appreciate the actionable feedback from the ACs and external reviewers. This work is partially supported by the National Science Foundation, under grant no. 2006854.

9 REFERENCES

[1] Julia Alexander. 2019. YouTube moderation bots punish videos tagged as 'gay' or 'lesbian,' study finds. The Verge. Retrieved from https://www.theverge.com/2019/9/30/20887614/youtube-moderation-lgbtq-demonetization-terms-words-nerd-city-investigation

[2] Ali Alkhatib and Michael Bernstein. 2019. Street–level algorithms: A theory at the gaps between policy and decisions. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), Association for Computing Machinery, New York, NY, USA, 1–13. DOI:https://doi.org/10.1145/3290605.3300760

[3] Carolina Are. 2021. The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. Fem Media Stud (2021). DOI:https://doi.org/10.1080/14680777.2021.1928259

[4] Anna Veronica Banchik. 2020. Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content. New Media Soc (March 2020), 146144482091272. DOI:https://doi.org/10.1177/1461444820912724

[5] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. Proc ACM Hum Comput Interact 5, CSCW1 (April 2021), 1–34. DOI:https://doi.org/10.1145/3449148

[6] Alex Barker and Hannah Murphy. 2020. YouTube reverts to human moderators in fight against misinformation. Financial Times. Retrieved from https://www.ft.com/content/e54737c5-8488-4e66-b087-d1ad426ac9fa

[7] Sophie Bishop. 2018. Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. Convergence: The International Journal of Research into New Media Technologies 24, 1 (2018), 69–84. DOI:https://doi.org/10.1177/1354856517736978

[8] Hannah Bloch-Wehba. 2020. Automation in Moderation. Cornell Int Law J (2020).

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qual Res Psychol 3, 2 (January 2006), 77–101. DOI:https://doi.org/10.1191/1478088706qp063oa

[10] Ragnhild Brøvig-Hanssen and Ellis Jones. 2021. Remix's retreat? Content moderation, copyright law and mashup music: New Media Soc (June 2021). DOI:https://doi.org/10.1177/14614448211026059

[11] Amy Bruckman, Pavel Curtis, Cliff Figallo, and Brenda Laurel. 1994. Approaches to managing deviant behavior in virtual communities. Association for Computing Machinery, New York, New York, USA. DOI:https://doi.org/10.1145/259963.260231

[12] Robyn Caplan and Tarleton Gillespie. 2020. Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. Soc Media Soc 6, 2 (2020). DOI:https://doi.org/10.1177/2056305120936636

[13] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 3213–3226.

[14] Stevie Chancellor, Jessica Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16, ACM Press, New York, New York, USA. DOI:https://doi.org/http://dx.doi.org/10.1145/2818048.2819963

[15] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. ACM Transactions on Computer-Human Interaction (TOCHI) 29, 4 (March 2022). DOI:https://doi.org/10.1145/3490499

[16] 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 (November 2017), 1–22. DOI:https://doi.org/10.1145/3134666

[17] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. Proc ACM Hum Comput Interact 2, CSCW (November 2018), 1–25. DOI:https://doi.org/10.1145/3274301

[18] George Christodoulides, Maximilian H.E.E. Gerrath, and Nikoletta T. Siamagka. 2021. Don't be rude! The effect of content moderation on consumer-brand forgiveness. Psychol Mark 38, 10 (October 2021), 1686–1699. DOI:https://doi.org/10.1002/MAR.21458

[19] Simon Copland. 2020. Reddit quarantined: Can changing platform affordances reduce hateful material online? Internet Policy Review 9, 4 (2020), 1–26. DOI:https://doi.org/10.14763/2020.4.1516

[20] Kelley Cotter. 2019. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. New Media Soc 21, 4 (April 2019), 895–913. DOI:https://doi.org/10.1177/1461444818815684

[21] Kelley Cotter. 2021. "Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. Inf Commun Soc (2021). DOI:https://doi.org/10.1080/1369118X.2021.1994624

[22] David Craig and Stuart Cunningham. 2019. Social media entertainment: The new intersection of Hollywood and Silicon Valley. NYU Press.

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

[24] Dipto Das, Carsten Østerlund, and Bryan Semaan. 2021. "Jol" or "Pani"?: How Does Governance Shape a Platform's Identity? Proc ACM Hum Comput Interact 5, CSCW2 (October 2021). DOI:https://doi.org/10.1145/3479860

[25] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic Transparency in the News Media. Digital Journalism 5, 7 (August 2017), 809–828. DOI:https://doi.org/10.1080/21670811.2016.1208053

[26] Tawanna R. Dillahunt, Xinyi Wang, Earnest Wheeler, Hao Fei Cheng, Brent Hecht, and Haiyi Zhu. 2017. The sharing economy in computing: A systematic literature review. Proc ACM Hum Comput Interact 1, CSCW (November 2017), 38. DOI:https://doi.org/10.1145/3134673

[27] Nicola Döring and M. Rohangis Mohseni. 2020. Gendered hate speech in YouTube and YouNow comments: Results of two content analyses. Studies in Communication and Media 9, 1 (March 2020), 62–88. DOI:https://doi.org/10.5771/2192-4007-2020-1-62

[28] Nicola Döring and M Rohangis Mohseni. 2019. Communication Research Reports Fail videos and related video comments on YouTube: a case of sexualization of women and gendered hate speech? Communication Research Reports 36, 3 (2019), 254–264. DOI:https://doi.org/10.1080/08824096.2019.1634533

[29] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. Conference on Human Factors in Computing Systems - Proceedings (April 2020). DOI:https://doi.org/10.1145/3313831.3376293

[30] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. Proc ACM Hum Comput Interact 4, CSCW1 (May 2020). DOI:https://doi.org/10.1145/3392845

[31] Casey Fiesler, Jialun Aaron Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit rules! Characterizing an ecosystem of governance. 12th International AAAI Conference on Web and Social Media, ICWSM 2018 (2018), 72–81.

[32] Padhraig S. Fleming, Jadbinder Seehra, Argy Polychronopoulou, Zbys Fedorowicz, and Nikolaos Pandis. 2013. A PRISMA assessment of the reporting quality of systematic reviews in orthodontics. Angle Orthod 83, 1 (January 2013), 158–163. DOI:https://doi.org/10.2319/032612-251.1

[33] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. New Media Soc 20, 12 (December 2018), 4492–4511. DOI:https://doi.org/10.1177/1461444818776611

[34] Anna Gibson. 2019. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. Soc Media Soc 5, 1 (January 2019), 205630511983258. DOI:https://doi.org/10.1177/2056305119832588

[35] Tarleton Gillespie. 2010. The politics of 'platforms.' New Media Soc 12, 3 (February 2010), 347–364. DOI:https://doi.org/10.1177/1461444809342738

[36] Tarleton Gillespie. 2018. Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press. Retrieved from https://www.degruyter.com/document/doi/10.12987/9780300235029/html

[37] Google. YouTube Community Guidelines enforcement. Retrieved from https://transparencyreport.google.com/youtube-policy/removals?hl=en

[38] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data Soc 7, 1 (January 2020), 205395171989794. DOI:https://doi.org/10.1177/2053951719897945

[39] James Grimmelmann. 2015. The Virtues of Moderation. Yale Journal of Law and Technology 17, (2015). Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/yjolt17&id=42&div=&collection=

[40] Helen Mayer Hacker. 1951. Women as a minority group. Social Forces 30, 1 (October 1951), 60–69. DOI:https://doi.org/10.2307/2571742

[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. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021). DOI:https://doi.org/10.1145/3479610

[42] Kai Hollander, Mark Colley, Enrico Rukzio, and Andreas Butz. 2021. A taxonomy of vulnerable road users for hci based on a systematic literature review. Conference on Human Factors in Computing Systems - Proceedings (May 2021). DOI:https://doi.org/10.1145/3411764.3445480

[43] 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 (October 2021). DOI:https://doi.org/10.1145/3476057

[44] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. Proc ACM Hum Comput Interact 3, CSCW (November 2019), 1–33. DOI:https://doi.org/10.1145/3359294

[45] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. ACM Transactions on Computer-Human Interaction 26, 5 (July 2019), 1–35. DOI:https://doi.org/10.1145/3338243

[46] 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 (October 2021), 30. DOI:https://doi.org/10.1145/3479525

[47] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter?: User behavior after content removal explanations on reddit. Proc ACM Hum Comput Interact 3, CSCW (2019). DOI:https://doi.org/10.1145/3359252

[48] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (2022). DOI:https://doi.org/10.1145/3491102.3517505

[49] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. ACM Transactions on Computer-Human Interaction 25, 2 (March 2018), 1–33. DOI:https://doi.org/10.1145/3185593

[50] Jialun Aaron Jiang. 2020. Identifying and addressing design and policy challenges in online content moderation. CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2020) (2020), 1–7. DOI:https://doi.org/10.1145/3334480.3375030

[51] 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 (November 2019). DOI:https://doi.org/10.1145/3359157

[52] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2023. A Trade-off-centered Framework of Content Moderation. ACM Trans. Comput.-Hum. Interact. (TOCHI) (2023). DOI:https://doi.org/10.1145/3534929

[53] 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 (August 2021). DOI:https://doi.org/10.1371/JOURNAL.PONE.0256762

[54] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in Reddit's moderation practices. Proc ACM Hum Comput Interact 4, GROUP (January 2020), 1–35. DOI:https://doi.org/10.1145/3375197

[55] D. Bondy Valdovinos Kaye and Joanne E. Gray. 2021. Copyright Gossip: Exploring Copyright Opinions, Theories, and Strategies on YouTube: Soc Media Soc 7, 3 (August 2021). DOI:https://doi.org/10.1177/20563051211036940

[56] Barbara Kitchenham. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering.

[57] Kate Klonick. 2017. The New Governors: The People, Rules, and Processes Governing Online Speech. Harv Law Rev 131, (2017). Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/hlr131&id=1626&div=73&collection=journals

[58] Susanne Kopf. 2022. Corporate censorship online: Vagueness and discursive imprecision in YouTube's advertiser-friendly content guidelines. New Media Soc (February 2022). DOI:https://doi.org/10.1177/14614448221077354/ASSET/IMAGES/LARGE/10.1177_14614448221077354-FIG1.JPEG

[59] Sarah Koslov. 2019. Incitement and the Geopolitical Influence of Facebook Content Moderation. Georgetown Law Technology Review 4, (2019). Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/gtltr4&id=195&div=9&collection=journals

[60] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021). DOI:https://doi.org/10.1145/3476075

[61] Yubo Kou and Xinning Gui. 2020. Mediating Community-AI Interaction through Situated Explanation: the Case of AI-Led Moderation. Proc ACM Hum Comput Interact 4, CSCW2 (October 2020), 1–27. DOI:https://doi.org/10.1145/3415173

[62] 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. Association for Computing Machinery, New York, NY, USA. Retrieved from https://doi.org/10.1145/3411764.3445279

[63] Robert E. Kraut, Paul Resnick, and Sara Kiesler. 2011. Building successful online communities: evidence-based social design.

[64] Cliff Lampe and Erik Johnston. 2005. Follow the (Slash) dot: Effects of Feedback on New Members in an Online Community. Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work - GROUP '05 (2005). DOI:https://doi.org/10.1145/1099203

[65] Kyle Langvardt. 2017. Regulating Online Content Moderation. Georgetown Law Journal 106, (2017). Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/glj106&id=1367&div=39&collection=journals

[66] Renkai Ma and Yubo Kou. 2021. "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. PACM on Human Computer Interaction 5, CSCW2 (2021), 1–26. DOI:https://doi.org/https://doi.org/10.1145/3479573

[67] Renkai Ma and Yubo Kou. 2022. "I am not a YouTuber who can make whatever video I want. I have to keep appeasing algorithms": Bureaucracy of Creator Moderation on YouTube. In Companion Computer Supported Co-operative Work and Social Computing (CSCW'22 Companion). Retrieved from https://doi.org/10.1145/3500868.3559445

[68] Renkai Ma and Yubo Kou. 2022. "I'm not sure what difference is between their content and mine, other than the person itself": A Study of Fairness Perception of Content Moderation on YouTube. Proc ACM Hum Comput Interact 6, CSCW2 (2022), 28. DOI:https://doi.org/10.1145/3555150

[69] Taniya Mapp. 2013. Understanding phenomenology: the lived experience. Br J Midwifery 16, 5 (September 2013), 308–311. DOI:https://doi.org/10.12968/BJOM.2008.16.5.29192

[70] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2010. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. International Journal of Surgery 8, 5 (January 2010), 336–341. DOI:https://doi.org/10.1016/J.IJSU.2010.02.007

[71] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and PRISMA Group. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Ann Intern Med (August 2009). Retrieved from https://sci-hub.se/https://doi.org/10.7326/0003-4819-151-4-200908180-00135

[72] Melissa J. Morgans. 2017. Freedom of Speech, the War on Terror, and What's YouTube Got to Do with It: American Censorship during Times of Military Conflict. Federal Communications Law Journal 69, (2017). Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/fedcom69&id=163&div=&collection=

[73] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media Soc 20, 11 (2018), 4366–4383. DOI:https://doi.org/10.1177/1461444818773059

[74] Sarah Myers West, Sigrid Kannengießer, and Sebastian Kubitschko. 2017. Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms. Media Commun 5, 3 (September 2017), 28–36. DOI:https://doi.org/10.17645/MAC.V5I3.989

[75] Chloe Nurik. 2019. "Men Are Scum": Self-Regulation, Hate Speech, and Gender-Based Censorship on Facebook. Int J Commun 13, (2019).

[76] Faiza Patel and Laura Hecht-Felella. 2021. Oversight Board's First Rulings Show Facebook's Rules Are a Mess. Just Security. Retrieved from https://www.justsecurity.org/74833/oversight-boards-first-rulings-show-facebooks-rules-are-a-mess/

[77] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work 13-16-November-2016, (November 2016), 369–374. DOI:https://doi.org/10.1145/2957276.2957297

[78] Jessica A Pater, Rachel Pfafman, Amanda Coupe, Chanda Phelan, Tammy Toscos, and Maia Jacobs. 2021. Standardizing Reporting of Participant Compensation in HCI: A Systematic Literature Review and Recommendations for the Field. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (2021). DOI:https://doi.org/10.1145/3411764

[79] Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. In Approaches to emotion. 197–219.

[80] Robert Plutchik and Henry Kellerman. 2013. Theories of Emotion. Academic Press.

[81] Ondřej Procházka. 2019. Making sense of facebook's content moderation: A posthumanist perspective on communicative competence and internet memes. Signs and Society 7, 3 (September 2019), 362–397. DOI:https://doi.org/10.1086/704763/ASSET/IMAGES/LARGE/FG3.JPEG

[82] Katie Reid, Paul Flowers, and Michael Larkin. 2005. Exploring lived experience. Psychologist 18, 1 (2005), 20–23.

[83] Sarah T. Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. Media Studies Publications (January 2016). Retrieved from https://ir.lib.uwo.ca/commpub/12

[84] Sarah T. Roberts. 2018. Digital detritus: "Error" and the logic of opacity in social media content moderation. First Monday 23, 3 (March 2018). DOI:https://doi.org/10.5210/fm.v23i3.8283

[85] Laura Savolainen. 2022. The shadow banning controversy: perceived governance and algorithmic folklore: Media Cult Soc (March 2022). DOI:https://doi.org/10.1177/01634437221077174

[86] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021), 33. DOI:https://doi.org/10.1145/3479512

[87] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (2017). DOI:https://doi.org/10.1145/3025453

[88] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Z. Tan, and Amy X. Zhang. 2021. Modular Politics: Toward a Governance Layer for Online Communities. Proc ACM Hum Comput Interact 5, CSCW1 (2021). DOI:https://doi.org/10.1145/3449090

[89] Sarita Schoenebeck, Carol F. Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice. Proc ACM Hum Comput Interact 5, CSCW1 (April 2021), 1–18. DOI:https://doi.org/10.1145/3449076

[90] Diana Secara. 2015. The Role of Social Networks in the Work of Terrorist Groups. Research and Science Today, 77–83.

[91] 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 (CSCW '17), Association for Computing Machinery, New York, NY, USA, 111–125. DOI:https://doi.org/10.1145/2998181.2998277

[92] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. New Media Soc 21, 7 (July 2019), 1417–1443. DOI:https://doi.org/10.1177/1461444818821316

[93] Spandana Singh. 2019. Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content. Retrieved from https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

[94] United Nations. 2018. Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression. United Nations Human Rights Office of The High Commissioner. Retrieved from https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx

[95] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. Proc ACM Hum Comput Interact 3, CSCW (November 2019), 163. DOI:https://doi.org/10.1145/3359265

[96] Stephen Stanford. 2018. YouTube and the Adpocalypse: How Have The New YouTube Advertising Friendly Guidelines Shaped Creator Participation and Audience Engagement?

[97] Janet. Sternberg. 2012. Misbehavior in Cyber Places : the Regulation of Online Conduct in Virtual Communities on the Internet. University Press of America.

[98] Elizabeth Stowell, Mercedes C. Lyson, Herman Saksono, Reneé C. Wurth, Holly Jimison, Misha Pavel, and Andrea G. Parker. 2018. Designing and evaluating mHealth interventions for vulnerable populations: A systematic review. Conference on Human Factors in Computing Systems - Proceedings 2018-April, (April 2018). DOI:https://doi.org/10.1145/3173574.3173589

[99] 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. Int J Commun 13, (2019). Retrieved from https://ijoc.org/index.php/ijoc/article/view/9736

[100] Jeanna Sybert. 2021. The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban: New Media Soc (February 2021). DOI:https://doi.org/10.1177/1461444821996715

[101] Katie Salen Tekinbaş, Krithika Jagannath, Ulrik Lyngs, and Petr Slovák. 2021. Designing for Youth-Centered Moderation and Community Governance in Minecraft. ACM Transactions on Computer-Human Interaction (TOCHI) 28, 4 (July 2021). DOI:https://doi.org/10.1145/3450290

[102] Pamela Bilo Thomas, Daniel Riehm, Maria Glenski, and Tim Weninger. 2021. Behavior Change in Response to Subreddit Bans and External Events. IEEE Trans Comput Soc Syst 8, 4 (August 2021), 809–818. DOI:https://doi.org/10.1109/TCSS.2021.3061957

[103] Amaury Trujillo and Stefano Cresci. 2022. Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald. (January 2022). DOI:https://doi.org/10.48550/arxiv.2201.06455

[104] Milo Z Trujillo, Samuel F Rosenblatt, Guillermo De, Anda Jáuregui, Emily Moog, Briane Paul, V Samson, Laurent Hébert-Dufresne, and Allison M Roth. 2021. When the Echo Chamber Shatters: Examining the Use of Community-Specific Language Post-Subreddit Ban. (June 2021). DOI:https://doi.org/10.48550/arxiv.2106.16207

[105] Rebecca Tushnet. 2019. Content Moderation in an Age of Extremes. Case Western Reserve Journal of Law, Technology and the Internet 10, (2019).

[106] Tom Tyler, Matt Katsaros, Tracey Meares, and Sudhir Venkatesh. 2021. Social media governance: can social media companies motivate voluntary rule following behavior among their users? J Exp Criminol 17, 1 (March 2021), 109–127. DOI:https://doi.org/10.1007/S11292-019-09392-Z/FIGURES/3

[107] Jirassaya Uttarapong and Rae Jereza. 2022. Social Support in Digital Patronage: OnlyFans Adult Content Creators as an Online Community; Social Support in Digital Patronage: OnlyFans Adult Content Creators as an Online Community. CHI Conference on Human Factors in Computing Systems Extended Abstracts (2022). DOI:https://doi.org/10.1145/3491101

[108] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. In Proceedings of the ACM on Human-Computer Interaction, Association for Computing Machinery, 1–22. DOI:https://doi.org/10.1145/3415238

[109] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021), 28. DOI:https://doi.org/10.1145/3476059

[110] Richard Ashby Wilson and Molly K. Land. 2021. Hate Speech on Social Media: Content Moderation in Context. Conn Law Rev (2021). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3690616

[111] Claes Wohlin. 2014. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14 (2014). DOI:https://doi.org/10.1145/2601248

[112] Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng (Polo) Chau, and Diyi Yang. 2021. RECAST: Enabling User Recourse and Interpretability of Toxicity Detection Models with Interactive Visualization. Proc ACM Hum Comput Interact 5, CSCW1 (April 2021), 1–26. DOI:https://doi.org/10.1145/3449280

[113] Yukun Yang. 2019. When power goes wild online: How did a voluntary moderator's abuse of power affect an online community? Proceedings of the Association for Information Science and Technology 56, 1 (January 2019), 504–508. DOI:https://doi.org/10.1002/PRA2.55

[114] Mustafa Mikdat Yildirim, Jonathan Nagler, Richard Bonneau, and Joshua A. Tucker. 2021. Short of Suspension: How Suspension Warnings Can Reduce Hate Speech on Twitter. Perspectives on Politics (2021), 1–13. DOI:https://doi.org/10.1017/S1537592721002589

[115] Jilian C. York and David Greene. 2020. How to put COVID-19 content moderation into context. TeachStream. Retrieved from https://www.brookings.edu/techstream/how-to-put-covid-19-content-moderation-into-context/

[116] YouTube. 2023. YouTube Partner Program overview & eligibility. YouTube Help. Retrieved from https://support.google.com/youtube/answer/72851?hl=en

[117] Jing Zeng and D. Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. Policy Internet 14, 1 (March 2022), 79–95. DOI:https://doi.org/10.1002/POI3.287

[118] Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised euphemism detection and identification for content moderation. Proc IEEE Symp Secur Priv 2021-May, (May 2021), 229–246. DOI:https://doi.org/10.1109/SP40001.2021.00075

[119] Jonathan Zittrain. 2019. A Jury of Random People Can Do Wonders for Facebook. The Atlantic. Retrieved from https://www.theatlantic.com/ideas/archive/2019/11/let-juries-review-facebook-ads/601996/

[120] Facebook Community Standards. Retrieved from https://transparency.fb.com/policies/community-standards/

[121] Instagram. 2023. Community Guidelines | Instagram Help Center. Retrieved from https://help.instagram.com/477434105621119

[122] Paperpile. 2019. The best academic research databases. Retrieved from https://paperpile.com/g/academic-research-databases/

Received July 2022, revised January 2023, accepted March 2023.