Opportunities, Tensions, and Challenges in Computational Approaches to Addressing Online Harassment

Evey Huang
Northwestern University
Evanston, Illinois, USA
eveyhuang@u.northwestern.edu

Abhraneel Sarma
Northwestern University
Evanston, Illinois, USA
abhraneel4@u.northwestern.edu

Sohyeon Hwang
Northwestern University
Evanston, Illinois, USA
sohyeonhwang@u.northwestern.edu

Eshwar Chandrasekharan
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA
eshwar@illinois.edu

Stevie Chancellor
University of Minnesota
Minneapolis, Minnesota, USA
steviec@umn.edu

ABSTRACT

Given the scale at which online harassment occurs, researchers and practitioners alike have turned to computationally driven approaches to address it. However, because harassment is highly contextual and personal, designing effective solutions to this problem can be extremely challenging. This paper examines how harassment-mitigation systems studied in human-computer interaction (HCI) consider victim-centered principles in their design. Through a scoping literature review and close reading of 17 papers, we contribute—(1) a characterization of how novel and existing systems consider victims' identity characteristics, definitions of harassment, and preferred strategies for dealing with harassment; (2) challenges faced by the systems along these dimensions to surface limitations, gaps, and tensions; (3) practical recommendations for researchers, designers, and practitioners to overcome these challenges. In doing so, we offer potential new directions to positively design computational approaches to addressing online harassment with victim-centered principles in mind.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**.

KEYWORDS

online harassment, moderation, computational tools

ACM Reference Format:
Evey Huang, Abhraneel Sarma, Sohyeon Hwang, Eshwar Chandrasekharan, and Stevie Chancellor. 2024. Opportunities, Tensions, and Challenges in Computational Approaches to Addressing Online Harassment. In *Designing Interactive Systems Conference (DIS '24), July 1–5, 2024, IT University of Copenhagen, Denmark.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3643834.3661623

1 INTRODUCTION

Online harassment is an ongoing challenge for online platforms. Formally defined, online harassment is targeted, interpersonal violence "where the victim believes that they have been harmed by one or more individuals" [69, 115]. In 2021, 41% of adults in the United States reported experiencing some form of online harassment [1]. Online harassment disproportionately affects marginalized groups, such as women, people of color, and LGBT+ people [33, 46, 98, 109]. From isolated instances of offensive name-calling to coordinated harassment campaigns such as #Gamergate, there is a pressing need to stop and address the negative consequences of online harassment [1, 51, 103, 109].

In response, researchers and technologists have turned to computational approaches to address harassment [40] and its disproportionate impacts [31, 56, 58, 104]. These are technological systems, tools, or affordances that scaffold and automate the work of handling online harassment prevention and mitigation [e.g., 19, 24]. For example, many Reddit moderators use AutoModerator, a tool that uses moderators' custom rules to automatically review and flag content [55]. These approaches offer scalable mechanisms through which users, moderators, and victims can address the harm done by perpetrators of online harassment.

State-of-the-art computational approaches must contend with two tensions in design and implementation: making headway on the challenges of scale and mitigation while also centering victims and their unique experiences with harassment. Victim-centered approaches for harassment have recently emerged because of user-centered design and moral imperatives to redress those harmed, especially when experiences are differential [11, 50, 92, 115]. Different personal *identity* characteristics change a potential victim's risk and vulnerability to harassment [44]. Qualitative research highlights how victims have varying *definitions* of harassment [12] and *preferences* for addressing it [92, 115]. Meanwhile, many factors contribute to errors in the identification and management of online harassment: technical challenges in automatically and accurately recognizing toxic interactions [21], narrow conceptualizations of what counts as abuse [11, 60], inability to parse cultural context in text [22, 46], and opaque 'blackbox' systems of moderation [15]. In short, online harassment is socially, culturally, and politically contingent in ways that require contextualization and care to combat and address the harm done to victims.

DIS '24, July 1–5, 2024, IT University of Copenhagen, Denmark

ACM ISBN 979-8-4007-0583-0/24/07.
https://doi.org/10.1145/3643834.3661623

Therefore, we ask: to what extent do computational research approaches to address online harassment align with victim-centered principles for redressing harm? To do so, we conduct a scoping literature review and close analysis of 17 research publications that build and evaluate computational approaches that address online harassment in human-computer interaction (HCI) research. Scoping reviews are focused literature reviews on newer and emergent topics in research. Instead of meta-review synthesis that is only possible through large corpus, scoping reviews in nascent research areas are powerful tools to understand trends, identify gaps, and clarify crucial concepts [3, 80], mirroring similar approaches in HCI and beyond [17, 61, 102]. Although concerns about online harassment are as old as the Internet, there is not yet consensus about how computational approaches can address those concerns—we believe anti-harassment technology innovation and research is happening rapidly and warrants further study and reflection, given its connection to design and praxis [97]. To ensure that victim-centered approaches that are important to design are materializing in our systems, the methods of a scoping review best afford the close consideration, evaluation and care required to achieve our research objectives [11, 92].

Drawing from prior work on online harassment [12, 44, 115] to focus on three key dimensions, we take an inductive, qualitative approach in analyzing our corpus of research publications to answer the following research questions:

- **RQ1.** How do computational approaches account for how *identity* traits affect experiences of harassment?
- **RQ2.** How do computational approaches incorporate users' *definitions* of harassment?
- **RQ3.** How do computational approaches incorporate users' *preferences* on how to cope and address harassment?

Our results show a growing trend toward computational systems adopting a victim-centered approach to addressing harm. Systems research acknowledges variable risks faced by users from marginalized groups, and *implicitly* accounts for identity characteristics during their design processes. However, translating findings about unique identity traits from formative studies to implementation is sometimes unclear. Due to the platform constraints of systems, we find a gulf between how victims report incidents and how they are classified (i.e., determined whether incidents are indeed cases of harassment); this leads to systems struggling to incorporate users' varying definitions of harassment. Promisingly, we also find examples of systems that explore opportunities to reduce this gulf [52, 56, 77]. Our analysis also highlights five key strategies for recourse post-harassment and how these strategies can effectively leverage community-building.

Building on our findings, we discuss how future work can center the *victim* as the key stakeholder and user of systems to handle harassment. We outline guiding principles to ground future systems design work, such as clearly stating who the intended users are, their needs, and how they relate to users' identities. Finally, we emphasize opportunities to develop technologies and tools that leverage social sharing, collaboration, and community to address harassment.

2 RELATED WORK

In this section, we provide an overview of online harassment and the distinct challenge the scale at which it occurs poses; the key layers of nuance that complicate online harassment as a domain to design computational systems for; and current approaches in place.

2.1 Online harassment and harm at scale

Online harassment remains a prevalent concern, with a 2021 Pew survey of U.S. adults revealing that 41% of survey respondents had personally experienced at least some form of online harassment [1]. While the survey focused on six forms of harassment, online harassment can cover many activities that result in harm to an individual [4, 32, 33, 76, 77], potentially leading to adverse behavior [70] within discussions. Following Xiao et al. [115], we draw on Krug et al. [69]'s broader definition of online harassment as targeted, interpersonal violence, "where the victim believes they have been harmed by one or more individuals" [115]. Although 'harassment' suggests repeated behaviors, our definition does not, as a single instance of harassment could be uniquely devastating to someone; we simply focus on interpersonal instances where individuals experience harm online. A broader definition gives space for the fact that different individuals might have different definitions of online harassment and harm, as described in §2.2.

Bad behaviors are not a uniquely online phenomenon, but prior work highlights how the global nature of the Internet and the scale and reach of interactions exacerbate them [40]. Indeed, the rate at which online harassment and harm occur leads to the crucial tension between scale and personal care of victims, which is a focal point of this work. The experience of online harassment can result in emotional distress and, in some extreme cases, self-harm for victims [109]. Because of scale, computational approaches have become critical in dealing with online harassment through automation [55, 62]. At the same time, the ways that technologies function at scale often mean they are inadequate for handling diverse and complex human interactions with care because they typically standardize as they take in inputs [41, 93, 107]. In the following sections, we describe how context and nuance are critical to adequately dealing with online harassment, highlighting why this tendency for standardization in scaling technologies is particularly troubling in this case (§2.2), and give an overview of current approaches for dealing with online harassment (§2.3).

2.2 Context and nuance in identifying and handling harassment

A critical concern with online harassment is how it may exacerbate existing inequalities and reflect societal prejudices. In particular, prior work highlights that individuals in minority groups disproportionately are victims of online harassment [26, 33, 91, 98, 105, 109]. Women, for example, are much likelier to experience sexual harassment online than men [1] because of toxic technocultures [78]; online communities centered around minority identities find themselves engaging in additional work to keep their spaces safe [26]. A qualitative study of Xbox Live gamers [44] lays out how Black female gamers are subjected to racialized sexism by other gamers, who linguistically profile them as Black and female through the voice feature of the gaming platform. Blackwell et al. [11] state that "abuse mitigation practices must ultimately protect and be informed by those who are most vulnerable, or the people who historically experience structural oppression." Thus, the kinds of characteristics about users that others glean online with harmful intent are of particular interest in this study. While *user characteristics* could cover a wide range of possible traits (e.g., someone who likes HCI), we focus on those that intersect with *identity* categories that are likely to make individuals more vulnerable to and marginalized by harassment, such as race, gender, age, religion, or nationality. This approach aligns with prior work from scholars like Jhaver et al. [56, 58], Blackwell et al. [10, 11, 12], and Schoenebeck et al. [92]. Therefore, our first research question focuses on how systems take into consideration these kinds of key user *characteristics*:

RQ1. How do computational approaches account for the ways that *identity* traits affect experiences of harassment?

The role of identity characteristics in the kind of harassment people experience highlights how cases of online harassment require nuance, context, and care to address. One challenge for computational approaches is the fact that individuals may have different perspectives and sensibilities about what constitutes online harassment. Similarly, different spaces may shape different interpretations of harassment, given their topic and scale. This is further complicated by dominant narratives of online harassment. Blackwell et al. [12] describe how some users "felt their harassment experiences did not fall within 'typical' expectations of what online harassment looks like, or even who is harassed online." As a result, victims may self-censor or downplay their experiences of online harassment. However, victims can benefit from sharing their experiences of online harassment when these experiences are validated by others [12]. Although it is critical for systems to recognize the diverse range of definitions of online harassment, the challenge with computational approaches is, unsurprisingly, sufficiently parsing and incorporating human and societal context and nuance [21, 22, 59, 60]. The vast amount of work that trains machine learning and deep learning models with large, annotated datasets to predict whether certain content is likely to be considered as harassment [e.g., 5, 16, 27, 81, 108, 113] can achieve high accuracy within this task by identifying large-scale patterns and phenomena of harm on different platforms. However, the reliance on algorithmic representations and abstractions of these approaches requires them to extract and compress the nuanced, complex contexts around individual users' experience of harm into quantitative data points [17]. Indeed, researchers in the growing sub-field of Fairness, Accountability, and Transparency in Machine Learning (FATML) and human centered machine learning (HCML) highlight how many computational approaches fail to parse and incorporate context sufficiently and instead perpetuate undesirable and harmful societal dynamics [2, 7, 17, 23, 46, 85, 111, 112]. As a result, our second research question focuses on *definitions*:

RQ2. How do computational approaches incorporate users' *definitions* of harassment?

However, even accounting for variations in identity characteristics and definitions, individuals may still have distinct preferences for how to cope with, and repair, the harm done [79, 87, 92, 109, 115]. For example, Schoenebeck et al. [92] suggest that different victims (of online harassment) may have different preferences for reparations and actions against the harassing online content. Similarly, a study examining adolescents' needs in addressing online harassment notes the importance of multiple stakeholders to support adolescent victims and lay out a variety of potential actions that victims felt were useful in addressing particular needs [115]. However, not every victim needed or wanted every action that others did. Given the multiple needs and potential paths to dealing with harassment noted by (potential) victims in research, our third research question focuses on these *user preferences* about how to address harassment:

RQ3. How do computational approaches incorporate users' *preferences* on how to cope and address harassment?

Through these research questions, we focus on three critical aspects that shape how online harassment plays out: (identity) characteristics that put people at variable levels of risk of harassment, definitions of harassment, and preferences for repair. In doing so, we center the perspective of the *victim* of online harassment as the "user" of these systems and the key stakeholders in our analyses. This is because when harassment does occur, sufficiently addressing the harassment means meeting the needs of victims.

2.3 Current responses to addressing online harassment

Current approaches and prior work focus on three main levels at which online harassment is dealt with: (1) the individual user, (2) communities and groups, and (3) platforms.[1] The individual user is the most basic unit that might act in relation to online harassment, a fact that is built into many social computing platforms. For example, Crawford and Gillespie [20] note the increasingly ubiquitous presence of the 'flagging' mechanism on social media platforms, which rely on individual users to report bad behavior and content. The 'flag' reflects a broader pattern of relying on end-users of platforms to report cases of harassment to platforms that host the interaction in the face of scalar concerns.

Community-driven approaches are similarly distributive but move beyond individuals to focus on groups. At this group level, addressing online harassment is often a coordinated effort. Prior work examines the massive amounts of labor volunteer teams perform to moderate content [45, 73, 95], the community-driven tools and strategies to combat anti-social behaviors [39, 55], and the dynamics of self-governance within and across those groups [9, 35, 49, 71]. While a great deal of this moderation work involves the basic act of reporting/removing content [45], group-level approaches to handling online harassment also involve deliberation [9], juried voting [29], and processes to handle conflict between individuals [63] and repairing harm [100].

A limitation of both the individual and community-based strategies are the affordances of the platform itself [6]. While not all digital spaces where online harassment occurs are platforms, many of the most prevalent spaces today are. Thus, individual and community-based strategies are often ultimately subject to critical decisions

[1] For an example of a multi-level model for thinking about online governance, see Jhaver et al. [57].

made by the platform [78]. The platform-level approach for addressing online harassment can primarily be seen in the algorithmic tools and filtering they deploy and the commercial content moderation workers they hire. However, as we have discussed, many automated tools fail to incorporate nuanced context and experience of harm and instead perpetuate undesirable and harmful societal dynamics [2, 7, 17, 22, 23, 46, 85, 111, 112]. They also fail to capture strategies individuals develop to subvert algorithms [18] or at times flag otherwise innocuous content and punishing users unfairly [99]. Meanwhile, investigations of commercial content moderation have raised serious ethical concerns about the traumatizing and exploitative nature of the work [89].

We see that as the phenomenon of harassment occurs at unfathomable scales [40], it draws in a complex ecosystem of social and technical actors at multiple levels of organization (individual, group, system). At each level, computational approaches and systems scaffold and automate human attempts to identify and handle online harassment, from reporting systems to end-user tools to platform algorithms. However, in doing so, scale *transforms* the problem of online harassment [107]: harassment becomes a series of flagged posts, reported comments, images, and cases to review, or filtered messages. In other words, an occurrence of harassment is flattened into a single instance like all others, which must be dealt with. Seaver [93] deftly describes how scale and care in the design of algorithms are often seen as in opposition with one another because of the uniformization of otherwise nuanced, contextual inputs.

However, Seaver [93] also argues that re-orienting the relationship between the two can generate new possibilities for our technological futures. In this work, we follow this call by evaluating existing computational tools proposed by researchers in light of the kind of care and nuance that is required to identify new opportunities for computational approaches to address online harassment and support victims of online harassment. To this end, **we focus on computationally-driven approaches that describe systems, tools, and other technological mechanisms that interface with *victims*** as we center them as users. While these approaches vary widely in their degree of technical sophistication, we note that this means we do not include a very large body of machine learning (ML) and natural language processing (NLP) work that seeks to improve the algorithmic accuracy in detection of harassment [see 60].

3 RESEARCH DESIGN

We conducted a scoping literature review to closely evaluate HCI publications that either present new computational approaches or evaluate existing ones that address online harassment. A scoping review [3, 28, 80, 82] is a literature review approach that produces a deep overview of previous research about a nascent domain, but refrains from judging the quality or weight of evidence provided by individual papers. Research in interventions for combating harassment are newer to HCI—given the emerging nature of this space, a scoping literature review is, therefore, an effective and ideal methodology to map out the existing strategies that have been adopted by researchers to address online harassment, identify gaps, and highlight directions for future research. Further, a scoping review allows us to integrate the sensitivity of victim-centered approaches into our analysis as a first-order concern while research is still developing and provide guidance while the area is still newer. Analytical approaches to reviews in HCI are common [102] and the ability to conduct closer reads on the conceptual decisions of research papers enables us to take a step back and identify opportunities for redirection.

Because different types of reviews serve different purposes, their outcomes also have different marks of rigor [43]. A common review type is a systematic literature review, which are evaluated based on their exhaustiveness and attempts to objectively synthesize a large body of research [43, 102]. HCI has many systematic reviews [102], including ones on health and online communities [37], reflection [8], and the sharing economy [25].[2] Scoping reviews, on the other hand, aim to "identify [the] nature and extent of research evidence" within a pre-determined scope [43]. In HCI, scoping reviews examine trends across research papers, such as ethics practices in SIGCHI [83] and human-drone interaction [47]. Although scoping reviews have no requirements for the size of the evidentiary basis, they do prioritize that the searching process be rigorous and analysis chosen to suit the research questions [3, 43]. Below we describe the scoping literature review process, overview the resulting data, and outline the analytical approach.

3.1 Scoping Literature Review Search

We next outline our processes for searching for potential manuscripts for inclusion. We follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping literature reviews (PRISMA-ScR) [75, 106] to identify and filter candidate papers.

3.1.1 Database and Search Boolean. We searched for research papers from the ACM Digital Library and the IEEE Digital Library, the two largest associations for HCI, computer-supported cooperated work, and computer science. Since the key element of our literature scope was **computationally-driven approaches that describe systems, tools, and other technological mechanisms that interface with *victims***, these two libraries with extensive coverage of HCI literature were deemed appropriate choices. Our most recent search was conducted on February 1st, 2024.

ACM Digital Library	Abstract: ("online") AND Abstract: ("harassment" "toxic" "moderation") AND ("system" "tool" "platform")
IEEE Digital Library	("Abstract":"online" AND ("Abstract":"toxic" OR "Abstract":"moderation" OR "Abstract":"harassment") AND ("Abstract": "system" OR "Abstract": "tool" OR "Abstract":"platform")

Table 1: The search booleans used in each library database to get an initial set of potential papers to include in the literature review.

Our search booleans were iteratively developed to focus on harm that happens online and results that include computational approaches, then expanding our search to include different types of online harm and approaches. Specifically, we first started by conducting a preliminary search for "online harassment system" within

[2] We highly recommend the recent work of Stefanidi et al. [102], which comprehensively covers reviews in HCI using, of course, a systematic review.

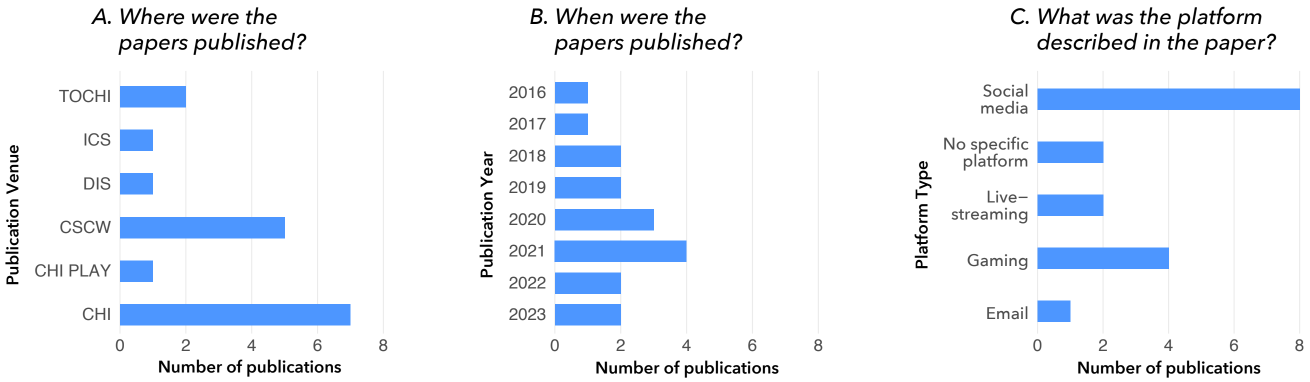


Figure 1: Details on publication venue, year of publication and online platform for the 17 papers that are included in our corpus

the titles of publications. With this search, we evaluated the relevance of results from the first page and iteratively added potential new keyword phrases, such as "toxic" and "tool", and expanded the search on the abstracts of publications. We then iteratively expanded and modified our boolean search term and carefully evaluated the relevance of the results by using a few articles which we considered good examples of the type of studies we wanted to conduct a scoping review of, until our search term returned all the seed publications. We excluded terms such as "model" and "detection" because the search results from these terms are mostly machine learning approaches that are out of the scope of this review because they do not interface with the *victims*. The final search booleans for each digital library were consistent and are shown in Table 1.

The search boolean for the ACM Digital Library resulted in 599 results, and 166 results from the IEEE Digital Library. After deduplicating and filtering the results to include peer-reviewed, full-scale archival journal and conference publications (eliminating workshops, non-archival posters, doctoral consortia, etc.), our initial set contained 561 full papers (401 from ACM and 160 from IEEE).

3.1.2 Filtering strategy and elimination. Next, we filtered this set of 561 papers by screening the full text. We include all studies that present, examine, and/or evaluate novel and existing computational approaches (technological systems, tools, designed affordances) for addressing online harassment. More specifically, we included studies that fulfilled the following criteria:

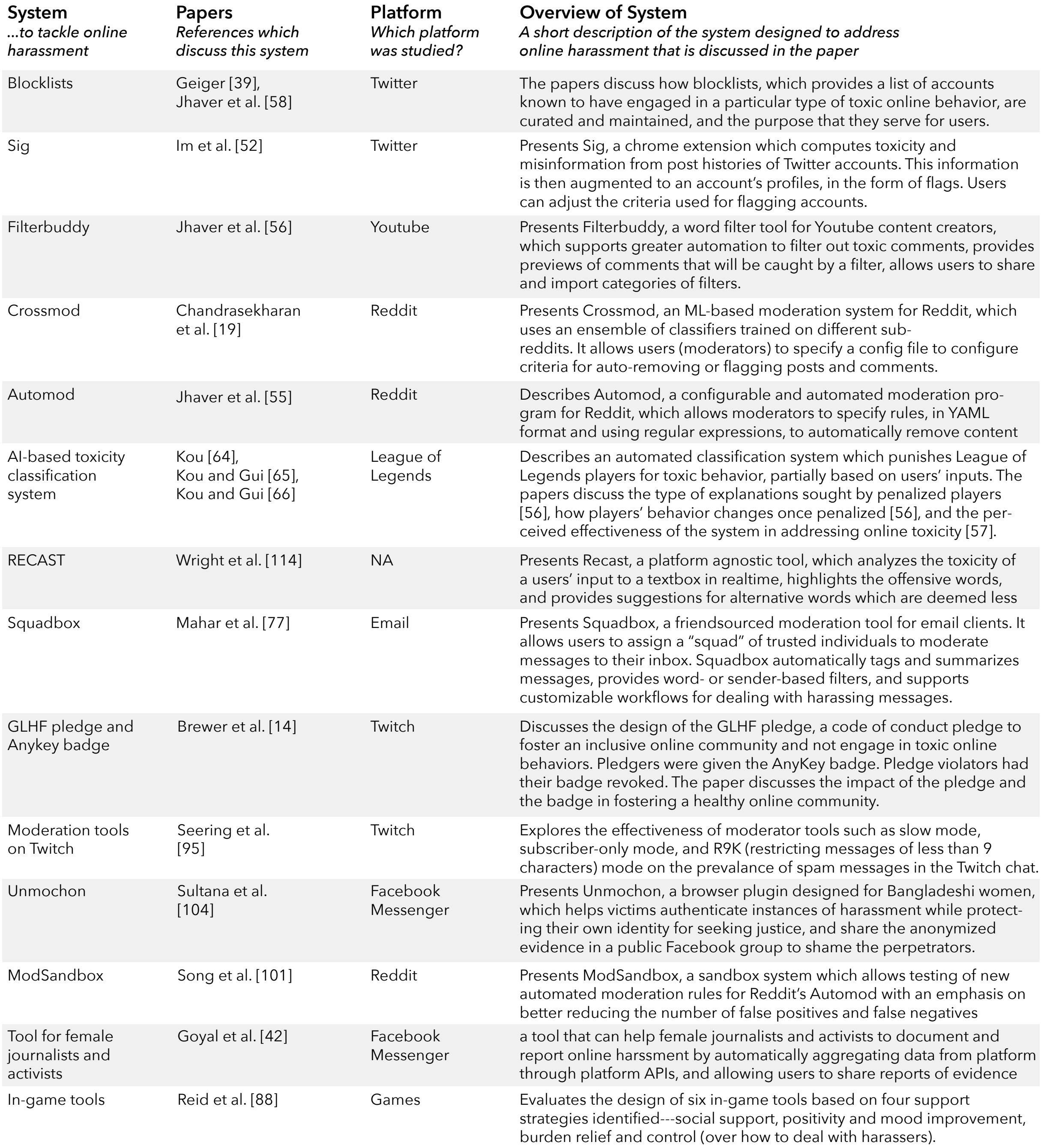
(1) the computational approach described should address harassment or toxic behavior on an online platform;
(2) the paper either describes the design and evaluation of a new computational approach OR provides a descriptive review or evaluation of an existing approach;
(3) the computational approach described should interface with a victim as the user; in other words, approaches that only detect toxic content or predict whether a post is going to be toxic or not were excluded;
(4) the approach that the paper describes should provide the users some form of recourse/intervention measures against the harassment

The first two criteria reflect the focus of this project on *computational approaches*, or various technological systems, tools, and designed affordances that are meant to address *online harassment*. Notably, we excluded papers that looked exclusively at the detection of online harassment, such as detecting toxicity or hate speech. A rich body of related work has focused on developing improved machine learning classifiers to better detect offensive or toxic content or content considered harassment [60], some of which has reached high accuracy [e.g., 5, 16, 27, 81, 108, 113]. However, when we attempted to include model development, we bifurcated our dataset and analysis because the papers were not comparable. Finally, our focus is on holistic processes of how computational systems *are designed and support* the end user who is a potential victim of online harassment. This is reflected in our third and fourth criteria, which focus on computational approaches that involve the victim as the user in addressing online harassment. After identifying papers to screen in, we conducted a manual backward citation search [30, 84] to identify additional relevant papers. We then applied our filtering strategy to these articles as well. After this round of elimination, we were left with 17 papers, which we used for the analysis presented in this study.

3.1.3 Data. To contextualize the findings, we provide a descriptive overview of the 17 papers in our corpus (Figure 1). All papers in our corpus have been published since 2016, suggesting a recent but growing interest in tackling online harassment through computational approaches that interface with victims-as-users. These papers have mostly been published in HCI venues such as CSCW (5) and CHI (7), and describe systems and approaches deployed on social media platforms like Twitter (3), gaming platforms like League of Legends (3), and online community platforms like Reddit (3). In one special case, the tool is not deployed on any particular platform but rather functions and was tested as a standalone tool that can be added to any major social computing platform [114]. Three of the systems are not deployed and evaluated on the existing platforms but only evaluated in experimental conditions [42, 88, 101]. Table 2 provides a brief overview of the system developed to combat online harassment, which is the paper's subject.

3.2 Analytical approach

Our review of work on online harassment emphasized the importance of centering victims, leading to our three research questions



System *...to tackle online harassment*	**Papers** *References which discuss this system*	**Platform** *Which platform was studied?*	**Overview of System** *A short description of the system designed to address online harassment that is discussed in the paper*
Blocklists	Geiger [39], Jhaver et al. [58]	Twitter	The papers discuss how blocklists, which provides a list of accounts known to have engaged in a particular type of toxic online behavior, are curated and maintained, and the purpose that they serve for users.
Sig	Im et al. [52]	Twitter	Presents Sig, a chrome extension which computes toxicity and misinformation from post histories of Twitter accounts. This information is then augmented to an account's profiles, in the form of flags. Users can adjust the criteria used for flagging accounts.
Filterbuddy	Jhaver et al. [56]	Youtube	Presents Filterbuddy, a word filter tool for Youtube content creators, which supports greater automation to filter out toxic comments, provides previews of comments that will be caught by a filter, allows users to share and import categories of filters.
Crossmod	Chandrasekharan et al. [19]	Reddit	Presents Crossmod, an ML-based moderation system for Reddit, which uses an ensemble of classifiers trained on different sub-reddits. It allows users (moderators) to specify a config file to configure criteria for auto-removing or flagging posts and comments.
Automod	Jhaver et al. [55]	Reddit	Describes Automod, a configurable and automated moderation program for Reddit, which allows moderators to specify rules, in YAML format and using regular expressions, to automatically remove content
AI-based toxicity classification system	Kou [64], Kou and Gui [65], Kou and Gui [66]	League of Legends	Describes an automated classification system which punishes League of Legends players for toxic behavior, partially based on users' inputs. The papers discuss the type of explanations sought by penalized players [56], how players' behavior changes once penalized [56], and the perceived effectiveness of the system in addressing online toxicity [57].
RECAST	Wright et al. [114]	NA	Presents Recast, a platform agnostic tool, which analyzes the toxicity of a users' input to a textbox in realtime, highlights the offensive words, and provides suggestions for alternative words which are deemed less
Squadbox	Mahar et al. [77]	Email	Presents Squadbox, a friendsourced moderation tool for email clients. It allows users to assign a "squad" of trusted individuals to moderate messages to their inbox. Squadbox automatically tags and summarizes messages, provides word- or sender-based filters, and supports customizable workflows for dealing with harassing messages.
GLHF pledge and Anykey badge	Brewer et al. [14]	Twitch	Discusses the design of the GLHF pledge, a code of conduct pledge to foster an inclusive online community and not engage in toxic online behaviors. Pledgers were given the AnyKey badge. Pledge violators had their badge revoked. The paper discusses the impact of the pledge and the badge in fostering a healthy online community.
Moderation tools on Twitch	Seering et al. [95]	Twitch	Explores the effectiveness of moderator tools such as slow mode, subscriber-only mode, and R9K (restricting messages of less than 9 characters) mode on the prevalance of spam messages in the Twitch chat.
Unmochon	Sultana et al. [104]	Facebook Messenger	Presents Unmochon, a browser plugin designed for Bangladeshi women, which helps victims authenticate instances of harassment while protecting their own identity for seeking justice, and share the anonymized evidence in a public Facebook group to shame the perpetrators.
ModSandbox	Song et al. [101]	Reddit	Presents ModSandbox, a sandbox system which allows testing of new automated moderation rules for Reddit's Automod with an emphasis on better reducing the number of false positives and false negatives
Tool for female journalists and activists	Goyal et al. [42]	Facebook Messenger	a tool that can help female journalists and activists to document and report online harssment by automatically aggregating data from platform through platform APIs, and allowing users to share reports of evidence
In-game tools	Reid et al. [88]	Games	Evaluates the design of six in-game tools based on four support strategies identified---social support, positivity and mood improvement, burden relief and control (over how to deal with harassers).

Table 2: Overview of the 17 papers that were included in our corpus. The system names in bold indicate that the papers present a newly developed system.

around user identity, user definitions, and user preferences. To answer our research questions, we collaboratively developed a rubric to guide our analysis and close reading of each paper. Two of the authors conducted a qualitative coding of the 17 research papers. During the qualitative analysis process, the two authors met with the rest of the research team to iteratively refine the rubric and analysis. We used an inductive thematic approach to analyze our data [13, 30]. Through iterative discussions among all authors, we analyzed and summarized the codes to develop our findings. Below, we describe how our close readings and qualitative analysis was operationalized for each RQ:

3.2.1 RQ1: User identity and personal traits. To answer RQ1, we developed two codes: (1A) does the design of the computational approach explicitly consider users' unique characteristics? (1B) what methods are used by the researchers and designers of the approach that enable them to consider users' unique characteristics? Because work on existing tools does not necessarily detail a design rationale or process, seven papers in our corpus presented novel systems or tools to evaluate how computational approaches take into account how identity or personal traits may impact how individuals experience online harassment.

3.2.2 *RQ2: User definitions of harassment.*

To answer RQ2, we created four codes initially to understand how computational approaches offer users *mechanisms of input* on what counts as harassment (2A,2B), and how these inputs are processed by the *mechanisms of classification* that label what is harassment (2C,2D): (2A) what is the process through which the systems receives input on whether a harassing or harmful action or interaction that has occurred? (2B) who is involved in providing this input? (2C) what is the process through which the system classifies whether an action or interaction is harassment? (2D) who is involved in performing this classification? We coded these by examining the description of the system presented in the paper.

This first step suggested that mechanisms of input and classification are often distinct and misaligned. Thus, we used the codes from the first step to develop two further codes: (2E) how aligned the two mechanisms (of input and classification) are in the sense that they converge or diverge on what counts as harassment; and (2F) how removed the two mechanisms are from one another, whether through transparency or communication of the agents involved. Together, these codes (2A-2F) structure insight into how computational approaches may fail to incorporate user definitions of what constitutes harassment.

3.2.3 *RQ3: Preferences for addressing harassment.*

To answer RQ3, we code the papers in our corpus along two dimensions: (3A) after an interaction is considered harassment (or in violation of norms of the online community), what action does the system take against the content and perpetrator/violator? and (3B) how is the victim involved in determining what action should be taken against the perpetrator? By mapping out and evaluating the range of currently used strategies, these codes help us answer RQ3—how computational approaches envision and constrain the ways users who have faced online harassment can play a role in how to seek recourse.

4 FINDINGS

In evaluating the corpus of 17 research papers, we underscore how current thinking about using computational systems (mis)align with the needs of victims that research on online harassment highlights.

4.1 How are the variable level of risks faced by different types of users taken into account?

Ten studies [14, 19, 42, 52, 56, 77, 88, 101, 104, 114] describe and evaluate *novel* computational approaches, and the remaining do not. Those that do not study novel approaches do not include a description of the design process and thus do not allow us to analyze whether the designers considered variable risks faced by users from different identities. Therefore, here we dive deeply into these ten studies.

Nine out of the ten studies specifically cited concerns of inclusion and safety and/or the disproportionate harms marginalized groups face in the motivating framing of their work (1A). One example of this is Brewer et al. [14], which presents a strong activist stance in the motivation for their work: "we align ourselves with other design activists who value an explicit orientation to social justice goals, place marginalized people at the center of design, and take a decided stance on the pressing issues of our day." Similarly, other studies [42, 56, 77, 88, 114] describe seeking to center the needs of minority groups who are often at significant risk for online harassment, echoing the call to action in Blackwell et al. [11]. To do so, with the exception of Brewer et al. [14], Reid et al. [88], all the studies in our corpus thoughtfully undertook formative qualitative studies and/or iteratively tested the designs of their novel tool; in contrast, Brewer et al. [14] was informed by prior community-engaged efforts (1B), and tools designed in Reid et al. [88] were inspired by previous research in the space Formative studies used interviews and open-ended surveys, which were methodologically and rhetorically central to identifying the needs of users who dealt with online harassment. This initial step enabled researcher-designers to refine the specifications of their computational approach. For example, Im et al. [52] conducted a smaller-scale pilot study of the system with 13 individuals, enabling the researchers to identify "issues that were crucial to cover for the final study" such as users' concerns about tool transparency. Similarly, Goyal et al. [42] conducted a focus group with 9 individuals, enabling researchers to have a deep understanding of these individuals' experience of online harassment before, during, and after it happens, and identify the lack of support for documenting evidence of the harm as the main challenge these individuals experience.

Despite this, we also observe that a given tool's implementation often does not clearly connect back to the stated motivations to center victims. The text of many studies did not explicitly discuss *what* or *how* the needs of marginalized individuals are met by their tool, even if the design itself was thoughtfully motivated by concerns of risks faced by marginalized groups (1B). In some cases, it was unclear from the description of the design rationale whether and how the characteristics of target groups were implemented into the system design. For instance, Mahar et al. [77] acknowledge that "certain groups such as young adults, women, and those who identify as LGBTQ" are more likely to face online harassment, and the participants they recruit for their formative study comprise primarily of women or non-binary individuals; yet how the needs determined through the formative study for the design of *Squadbox* relate to the identities of these individuals are not clearly outlined. Similar work in HCI/CSCW has found this misalignment between the self-stated intentions of researchers and the actualization of these commitments in methodology [17].

On the other hand, a positive example is Sultana et al. [104], which used a combination of 91 survey responses and 43 in-depth interviews with Bangladeshi women. These interviews led to *Unmochon*, a tool to help combat online sexual harassment on Facebook Messenger. The extensive work done in the pre-design stage makes clear how *Unmochon* is rooted in the needs, concerns, and preferences of women in Bangladesh who experience harassment. *Unmochon*'s design to specifically combat sexual harassment is directly connected to survey responses reporting unwanted and threatening messages that were distinctly gendered (e.g., requesting nudes, repetitively asking for marriage, threatening sexual violence, and so on). Similarly, Goyal et al. [42] uses a combination of focus groups and in-depth interviews with female journalists and activists to understand their experience of online harassment and how their occupational demands leaves them particularly vulnerable. They then designed a tool to specifically address those needs—to document and report the evidence of harm. Features of this tool are directly linked and connected to insights the researchers have learned in the early interview phase.

In contrast to the approaches adopted by Sultana et al. [104] and Goyal et al. [42], many studies design computational approaches envisioned to be used *generally* while also attempting to center the needs of the most vulnerable in shaping their design. Consequently, a disconnect between the motivation and execution through the design pipeline may emerge due to imprecisions or ambiguities about how specific needs or risks are addressed by a designed approach when presented as a general solution yet motivated by more specific concerns. This disconnect may be due to how explicitly the authors acknowledge the variable risks different individuals face throughout the design process. However, it is possible for generally-scoped computational approaches to connect back to the initial motivations of the disproportionate risk online harassment poses to marginalized communities. For example, *FilterBuddy* [56], a word filtering tool for YouTube content creators, seeks to center the needs of minority groups by intentionally oversampling for gender, racial, and sexual minorities in the formative interview study. The evaluation and discussion of *FilterBuddy* are connected to their original motivating goal, reporting participants' remarks on the potential for *FilterBuddy* to specifically aid users in marginalized groups to collaboratively customize stronger filtering protections.

Our analysis suggests that initial motivations to prioritize the needs of diverse and marginalized users, especially when implicit, may not always extend to actually *designing* for those users. Without answers to these questions, we argue here that researchers and designers may risk inadvertently producing designs that only produce a facade of addressing the needs of the most vulnerable, despite what we see as best intentions and thoughtful pre-design work.

4.2 How are users' potentially different definitions of harassment taken into account?

Next, we describe how users' differing definitions of harassment play out during system interactions. As most platforms operate under vague, broad definitions of harassment, considering users' potentially different definitions usually requires active user involvement. Sixteen of the papers in our corpus (except [114]) describe the *mechanisms of input*—the processes through which users who experience harassment can report incidents (2A, 2B)—and the *mechanisms of classification*—the processes through which these reports are validated (2C, 2D). In most cases, the *input* regarding instances of harassment is provided by the user of the tool. This mechanism is enacted for users in several ways, ranging from "flagging" in games [64–66, 88], to nominating moderators and providing them with instructions in *Squadbox* [77], to directly or indirectly configuring an automated system which classifies content as harassment [19, 52, 55, 56, 101]. Providing such *mechanisms of input* is the first step that platforms can take to support users with variable risks and needs; users and victims have different notions of harassment, and therefore are likely to report different content or interactions as such. The *mechanisms of classification* involves determining whether reported cases were indeed harassment. While input regarding the occurrence of harassment is usually reported at the individual user level, classification may happen at the user level [19, 42, 52, 55, 56, 77, 88, 104], the subcommunity level [14, 38, 58, 101], or at the overall platform level [64–66]. According to the multi-level model of online governance by Jhaver et al. [57], these levels map to different levels of power in online spaces.

A common approach to consider different definitions of harassment for different users is by having processes to influence or configure the mechanism of *classification* [19, 52, 55, 56, 77]. Providing users control and configurability over what an automated system classifies as harassment provides users with the ability to moderate their online content based on their own subjective and variable risk. Control and configurability also introduce transparency by allowing the user to explore the results of different configurations. For example, *Filterbuddy* [56], the word-filtering tool for Youtube, provides curated categories of word-filters based on identity attacks such as racist or sexist words that can be imported by the users. Users can manually create their own word-filter category, supporting user configuration of the automated classification system. Similarly, *Sig* [52], a tool that highlights Twitter profiles if they have a history of tweeting toxic content or misinformation, allows the user to set thresholds for the frequency of such tweets; profiles whose tweet history exceeds these thresholds will be flagged. Supporting users through configurability is perhaps taken to its most extreme by Mahar et al. [77] who design *Squadbox* using the principle "everything should be an option." These tools find the right balance of automated classification to handle challenges of scale, providing user agency over how the automated classification system makes decisions and transparency over the decision-making system. These approaches can shift the *mechanism of classification* from the top-level to the user or a sub-community level, reducing the distance between the two mechanisms (2E), and making the two mechanisms closely aligned insofar as sharing similar notions of harassment (2F). However, such configurable systems should be designed with care and caution as offering configurability can end up being a method that facilitates *generalizability* instead of a method that centers the needs of the most vulnerable. Here, we note that *FilterBuddy*'s strategy of offering customization to end users is common in other 'general' computational approaches (such as Chandrasekharan et al. [19], Mahar et al. [77]) seeking to meet the diverse needs of potentially very different users. *FilterBuddy*'s success in fulfilling its motivating framing comes through attending to how identity shapes experiences of harassment and reflections on how their work connects to this point—rather than from the fact that *FilterBuddy* itself offers user configuration or customization.

There were also examples of approaches other than configurability for successfully incorporating subjective definitions' of harassment. Take, for instance, the work of Brewer et al. [14], who allow the GLHF pledge-takers to report perceived violators and provide opinions on how the moderators ought to react through an online form. Many of these reports provided necessary context and nuanced opinions on how the offender should be punished. As another example, Twitter blocklists [38, 58] allow users to subscribe and block curated lists of Twitter accounts that have engaged in specific types of toxic behaviors online. Blocklists extend a primitive mechanism of *input* and *classification* on Twitter to one that can be deployed at scale as long as the user can find blocklists that combat a specific type of harassment. We argue that is only possible because the users and the moderators have a strong and shared understanding of harassment—in other words, strong alignment between the two mechanisms of *input* and *classification* (2F).

In systems where the two mechanisms are distant, both in terms of the levels of online governance and transparency, this situation may lead to users being distrustful of automated moderation approaches. This is evident in *League of Legends* [64–66], where users (players) can report harassment through an in-game *flagging function* that sends reports to a platform-level automated moderation system. Although players receive a notification if action was taken against the offender based on their report, such punishment appears rare. Most players are left uncertain whether their flags have an effect [66]. We argue that this lack of influence on and visibility into the *classification* process led to League of Legends players not trusting flagging as a *mechanism of input* for reporting harassment. Instead, players misappropriated the flag to perpetuate a culture of toxic meritocracy [66].

4.3 How are users involved in deciding how to address harassment?

In our corpus, we found five distinct approaches for addressing instances of harassment (3A): first, punitive strategies of moderation that punished or sanctioned bad behavior were the most common strategies for addressing harassment, including removal or message limiting [19, 55, 56, 88, 94], banning [65, 94], or publicly shaming perpetrators [104]; second, distancing strategies created space between victims and (past or potential) perpetrators by constructing barriers and partitioning online spaces. For example, labels marking users as toxic in Sig [52], the GLHF pledge badge [14], or blocklists on Twitter automating blocking of content from those included on the list [38, 58], and the addition of moderation intermediaries in Squadbox [77] jam the flow of content between potential victims and perpetrators; third, educational strategies for perpetrators were another strategy for addressing harassment towards more pro-social behaviors [14, 88, 94, 114]. For example, gamers who experience online harassment can use tools designed by Reid et al. [88] to send a message to the perpetrators to notify them of their behaviors. Similarly, community moderators can provide a guideline for the perpetrators on how to regain their status after their accounts are banned [14]. The fourth strategy is help the victims collect and report the evidence of harm to authorities [42, 104]. For instance, the tool designed by Goyal et al. [42] can help female journalists and activists who experience online harassment to automatically aggregate and collect evidence of online harassment across different social media platforms, and directly share this evidence with authorities. The final strategy is to provide emotional or social support to the victims [88], which includes showing them cute animal pictures, friendly messages from other players in the game, or a voice-line to hear positive voices from the in-game character.

Similar to our findings in §4.2, formative pre-design studies with users often directly drew from users' preferences for addressing harassment to appropriate design mechanisms (3B). Still, users are constrained by each system offering just one means of recourse. That is, users can decide how to address harassment *if* the strategy is part of the computational approach's design. However, our corpus suggests that the strategies across studies became meaningful ways of addressing harassment for users because of the surrounding dynamics of collaboration and community building. Approaches taking a punitive strategy both operated within community spaces and generated collaboration among individual users: after deploying *FilterBuddy* [56], the filtering tool for video comment moderation among content creators on YouTube, participants "were keen to engage in community building with other creators by leveraging the sharing feature" and expressed interesting collaborating with other users in their filtering work. Users of *Unmochon* Sultana et al. [104] similarly saw their public shaming of their harassers as an act of warning others, specifically shaming within an online community space (Facebook Group). Strategic creation of space between victims and perpetrators often emerged from existing relationships and communities, such as the construction and deployment of blocklists, which are "embedded in and emerge from counterpublic communities" [38], or the friends-as-moderators model for creating a protective barrier for victims in *Squadbox* [77]. Meanwhile, the strategy of educating perpetrators and providing social support was tightly tied to notions of pro-social efforts to build community (sometimes in a more literal sense of online communities). For example, in Brewer et al. [14], authors found that by adopting an "empathetic and educational stance towards even the most serious offenders", many ex-harassers followed the feedback they received about their behavior to get their badge back and reintegrate themselves into the "safe space" communities.

Together, the role of community in making strategies more meaningful ways to address harassment emphasizes the social nature of online harassment that makes purely technological solutions inadequate. The papers describe how community-building approaches to addressing harassment support the capacity for fostering collective action that can produce meaningful tools for users (such as the creation of tools such as blocklists [38, 56] and sharing of filtering strategies in [56]) and for social support and creating safe spaces (such as via the GLHF pledge [14] or the *Unmochon* community [104]). Alone, each strategy (punitive, space, educational) clearly articulates how an individual case of online harassment might be dealt with. However, the various ways that the community is implicated in, emerges from, and transforms how these strategies become useful points to a need to move away from thinking of cases of online harassment as individualized. In turn, it suggests that such communal spaces are arenas where users can build out sociotechnical interventions for how they want to see harassment addressed.

5 DISCUSSION

Together, our findings underscore current and nascent patterns in designing computational approaches to address online harassment.

Research Questions	Findings	Design Recommendations
RQ1: How are the variable level of risks faced by different types of users taken into account? (1A) Does the design of the computational approach explicitly consider users' unique characteristics? (1B) What methods are used by the researchers and designers of the approach that enable them to consider users' unique characteristics?	1A: Most systems specifically cited concerns of inclusion, safety, and disproportionate harms faced by marginalized groups in motivating their work. 1B: Many studies that followed user-centered design principles did not explicitly discuss what/how the variable level of risks faced by marginalized groups are met by the proposed systems.	(1) Clearly state who the intended users are, their needs, and how these are related to the users' identities. Clearly describe how the designed solutions addresses those stated needs of marginalized groups. (2) Clearly define how the success of designed systems can be measured as desired outcomes may vary and that diverse users may face new or unique challenges in effectively using such systems.
RQ2: How are users' potentially different definitions of harassment taken into account? (2A) What is the process through which the systems receives input on harassment? (2B) Who is involved in providing this input? (2C) What is the process through which the system classifies whether an action or interaction is harassment? (2D) Who is involved in performing this classification?	2A, 2B: Includes mechanisms of input ranging from ``flagging'' [66, 64, 65, 88], nominating moderators and providing them with instructions in [77], directly or indirectly configuring an automated classfier [19, 52, 55, 56, 101]. 2C, 2D: The mechanism of classification occurs at the user level [19, 42, 52, 55, 56, 77, 88, 104], the subcommunity level [14, 39, 58, 101], or at the overall platform level [64, 65, 66], mapping to three levels of governance [57].	(1) Reduce disconnect between mechanisms of input and classfiation at the platform level. For e.g., offering users explicit configurability, such as setting acceptable levels of toxicity or defining what kind of content should be filtered [52, 56]. (2) Enable the capacity for and create designs that shift who is engaged in classification, and to what extent, beyond the platform owners themselves. (3) Offer transparency ito users inputs. e.g.: status of reports, details on how reports are classified, explanations for decisions, and actions taken.
RQ3: How are users involved in deciding how to address harassment? (3A) After an interaction is considered harassment (or in violation of norms of the online community), what action does the system take against the content and perpetrator/violator? (3B) How is the victim involved in determining what action should be taken against the perpetrator?	3A: Five distinct approaches for addressing instances of harassment: (1) Punitive approaches [19, 55, 56, 65, 88, 95, 104]; (2) Distancing strategies to create space between victims and perpetrators [14, 39, 52, 58, 77]; (3) Educational strategies to move perpetrators towards more pro-social behaviors [14, 88, 95, 114]; (4) Help the victims collect and report the evidence of harm to authorities [42, 104]; (5) Provide emotional or social support to the victims [88] 3B: Users are constrained by one means of recourse that each system offers, even though those means of recourse are informed by users' preferences in formative studies.	(1) Offer alternative approaches for recoures, beyond punitive approaches to address harassment. For example, to offer educational opportunities to the offenders [14] and provide social and emotional support to the victims [88]. (2) Technological solutions alone may not be sufficient. Platforms and systems should build capacity for social sharing, collaboration, and community to address harassment. (3) Designers and researchers should adopt participatory design approaches to identify and explore opportunities for incorporating different options for addressing harm outside of the punitive into their systems.

Table 3: Summary of findings and design recommendations

In doing so, we highlight misalignments in sufficiently meeting the needs of victims along three key dimensions (identity, definition, and action), as well as ongoing attempts to address them. In the following sections, we discuss challenges and opportunities in computationally addressing online harassment from the tensions apparent in each of the core findings.

5.1 Make explicit who and what identities are being designed for

Prior work argues that designers of tools to combat online harassment should explicitly consider how certain users' identities might impact their online experience [11, 55, 56, 92, 109], specifically making calls to center the needs of the most vulnerable [11]. Our analysis reveals that while system designers adopted user-centered design practices to identify user needs, they often did not distinguish between the issues the broader public faces and those unique to marginalized users who face a greater risk of online harassment as their design work proceeded. While studies frequently cited the needs of marginalized groups as *motivating* their work, this primarily manifested in a pool of participants for the formative study that was relatively gender-balanced and gender-inclusive. Beyond this step, how the resulting tool related back to the specific needs and risks of these groups was frequently left unsaid.

In contrast, three studies in our corpus showcased the value of explicitly looping downstream decisions, analyses, and discussion back to the motivating goal of centering the needs of the most vulnerable: *Unmochon*, which completely focused on a specific marginalized group to clearly delineate their needs and focus on those needs throughout the design process; *Filterbuddy*, which intentionally over-sampled users from marginalized groups to guide their design work; and a tool Goyal et al. [42] designed to specifically support female journalists' and activists' needs. All three papers **clearly state who the intended users are, their needs, and how these are related to the users' identities**. Consequently, these systems explicitly described how their solution addresses the specific needs of marginalized groups, thereby articulating the values embodied by the technologies. Work in the area of critical computing—such as reflective design [96] and value-sensitive design [36]—has long shown how accounting for the values embedded into technologies is critical to understanding their outcomes and effects as well as opening up new design possibilities. Similarly, we call on future work addressing online harassment to clarify how their resulting system relates to the core motivating value underpinning the design work: centering the needs of the most vulnerable. That is, future work should address how their designed tool will impact various marginalized groups that are at greater risk of online harassment, and connect how the features of the system address the needs of these user groups. In doing so, such systems can concretely advance our understanding of designing computational tools to address harassment with these groups in mind [11, 91], as opposed to reverting to a broader and less-grounded sense of what harassment entails.

We noted in our findings that the one reason for the disconnect between motivation and execution may be due to the manuscript writing process rather than the design process itself, in which case the solution is for researchers to articulate these connections in the manuscript explicitly. However, we also recognize that other ethical questions underpin and may contribute to this disconnect. For example, histories of extractive research and design work that ultimately negatively impact communities may very reasonably disincline communities from working closely with researchers and designers [34, 86, 90, 110]. Moreover, researchers must remain cognizant of the labor, agency, and work that the community who care invest by participating in studies [48, 74, 91].

One promising direction our corpus points to is in offering users explicit configurability, such as setting acceptable levels of toxicity or defining what kind of content should be filtered [e.g. 52, 56] to address online harassment. Configurability may allow users to adopt the tool in ways that designers have not thought about, implicitly accommodating different notions of harassment and the variable needs of different identity groups. Because it shifts a crucial point of decision-making to users, configurability in theory offers the promise of user agency. However, current realizations of configurable systems, such as personalized content moderation tools [53, 54], can leave users confused as to what they are configuring and how that will impact the content shown on their feed [59]. We also observed that configurability alone was not sufficient to address the needs of the most vulnerable. As a result, there is much room for researchers and designers to improve the user experience of configurable systems. As part of this, we note that there is also much room to critically evaluate how users coming from different backgrounds might variably experience or be impacted by configurability. To this end, systems researchers and designers will need to clearly define how the success of these systems can be measured, considering that desired outcomes may vary and that diverse users may face new or unique challenges in effectively using such systems.

5.2 Reduce the gulf between identifying and classifying harassment

Our analysis underscores a disconnect between the users who experience and report incidents of harassment online and the actual mechanisms that process and validate reports of harassment as harmful cases to address, generally at the platform level. This gulf between users and platforms as mechanisms of input (identifying incidents) and of classification (deciding if incidents are indeed cases of harassment) creates many paths through which context, nuance, and understanding can be lost—in particular, through misalignments between what users and platforms understand as harassment. Prior work emphasizes how such disconnects are, ultimately, to the detriment of users experiencing harassment when platform decision-making is obscure and input mechanisms are thin [20, 57]. For instance, Crawford and Gillespie [20] highlights how ill-implemented flagging or reporting mechanisms can serve to protect the online platform from scrutiny over inaction towards addressing online harassment on the platform itself.

Here, we note that a core challenge in actually closing this gulf is the simple reality that the platforms that many tools are designed to address harassment on are closed systems, owned by large tech companies; many independent designers and researchers do not have a direct say in the design of these systems, and some systems do not offer APIs that would enable the building of tools to run alongside them. Our work underscores that platforms should open up these possibilities through APIs or enabling compatibility with external tools by other means. Repeatedly, the studies in our corpus emphasize the benefits of **enabling capacity for and creating designs that shift who is engaged in classification, and to what extent**, beyond the platform owners themselves. Systems such as *Sig* and *Crossmod*, for example, introduce new input mechanisms by having users specify thresholds or tune classifiers to proactively catch and filter content—leveraging the APIs of Twitter and Reddit, respectively. Meanwhile, systems such as Squadbox [77], Twitter blocklists [38], and Filterbuddy [56] introduce new mechanisms of classification through community-level sharing and collaboration to build well-tuned parameters and lists of what counts as harassing content. In each of these systems, the common pattern of identification-then-classification is flipped: classification becomes identification, where users and communities—not the platform—define the classifications of harassment and the system then uses their definition to accordingly identify content as harassing or not.

In light of this, researchers and designers may consider focusing on the development of standalone tools that do not rely on the platform's API but still continue expanding on the model of classification-as-identification noted above. For example, future work might consider building extensions for web browsers more generally as a point of entry to tackle issues of harassment on particular platforms that individuals use. Meanwhile, many potential paths can be taken at the platform level similarly. For example, platforms might make platform-level algorithmic definitions of harassment *responsive* to claims from users and communities of what constitutes harassment, with safeguards in place for inputs from trolls and spammers. Given that identification of harassment is already heavily distributed to users, platforms should also create interfaces that offer more insight to users about the status of cases of harassment they have reported, such as visual status update tools, better support for users about how cases of harassment are in fact classified and determined, and explanations for decisions about their cases. Beyond increasing user agency, reducing the gulf between the points of identifying and classifying harassment can increase users' trust in the system through transparency, expectation-setting, and offering a sense of control.

5.3 Provide clear and rich signals of the value of user input

Online systems rely on user participation in order to manage and moderate cases of online harassment because of the scale advantages of distributing this labor [20, 40, 71, 72]. In our corpus, almost all systems require user participation in one form or another to address harassment, and we consistently found that users valued knowing whether their participation made an impact or how their input was received [e.g. 14]. This highlights the importance of **providing clear and rich signals of the value of user inputs from the system** in order to foster an engaged and healthy online community.

While we observe many systems in our corpus that elicit user input in different forms, not all provide detailed feedback to the user about how their input will affect the system. This is evident from cases in our corpus where distrust grows in the absence of such feedback and affirmation of user inputs, as is clearly the case in League of Legends [64–66]. When users who provide input reporting on incidences of harassment are left unclear on the impact of their input, they may lose motivation to report instances of harassment, leave the community, or misappropriate the input tool itself.

A broader body of literature focused on building successful online communities has highlighted the advantages of creating incentives and rewards of desirable contributions [67, 68]. Thus, in noting that signals to users about their inputs matter, our work follows prior work in advocating for clear communication to users about how and why their participation matters for building successful sociotechnical systems. Future work should develop designs that transparently communicate, with precision, to users about how their inputs are used—or not—in systems and are being acted upon (or provide information on why no action was taken). A strong example of doing so in our corpus is *Filterbuddy* [56], which gives previews of comments that would be removed if a certain phrase is added as a filter, providing the user real-time feedback about how the system functions. In a similar vein, we note that future systems can leverage preview functions or other methods of showing counterfactuals of how different configurations of the system may look to provide a rich signal to the user of how their inputs affect the system, whether the system is able to achieve their goals, and when the system might fail to meet their expectations or needs.

5.4 Explore different options for addressing harassment

Most justice-seeking approaches employed by platforms involve punitive measures such as banning the accounts of or removing content posted by the perpetrators of an offense. Leaving aside the effectiveness of such measures, recent work has identified that social media users may prefer alternative approaches for recourse that are currently not supported by online platforms Schoenebeck et al. [92]. Yet, online platforms seem ill-equipped to transition away from these traditional approaches to justice as currently, there exists limited affordances within platforms to design for most such alternative approaches of seeking justice.

Our analysis reveals examples of systems successfully adopting alternative approaches to address harassment. *Unmochon* allows victims to publicly shame offenders by posting their offensive messages to a public Facebook group. While public shaming of perpetrators is punitive in nature, the public Facebook group used in *Unmochon* functions as a community space to air grievances and warn others about the perpetrators — victims of harassment find this method to be desirable, fair, and just [92]. On the other hand, Brewer et al. [14] discuss adopting an educational stance towards users whose AnyKey badge has been revoked for violating the GLHF pledge by engaging in offensive online behavior and providing offenders a pathway for redemption. In another example, Reid et al. [88] designed a set of tools that offer victims emotional and social support, such as receiving cute animal pictures and receiving friendly messages from other players. While all of these examples employ starkly different strategies for enacting justice, they are bound by a common thread of being grounded within the community and embody values that are shared by the community.

Given the social nature of harassment, our analysis suggests that technological solutions alone may not be sufficient. In many cases, the studies instead emphasize the importance of communal approaches to dealing with harassment. In light of these observations, we reiterate prior work [92] and point towards potential opportunities in exploring alternative justice strategies for online moderation systems. Additionally, we believe that **developing technologies and tools that build capacity for social sharing, collaboration, and community** remains an exciting opportunity for computationally addressing harassment. Even tools such as Squadbox, Filterbuddy, and blocklists, which employ punitive recourse mechanisms, are embedded in social groups. Jhaver et al. [56], for example, noted that in using the YouTube comment filtering tool, users both began to build community with one another outside of the technological limits of the tool itself and expressed enthusiasm for collaborating with other users to improve filtering strategies. We speculate whether participatory design approaches may be able to help researchers identify and explore opportunities for incorporating non-punitive options for addressing harm into their systems.

5.5 Limitations and Future Work

Like in all reviews that use search terms, it is possible that our keywords did not return results for relevant studies. Although we both experimented with keywords and conducted a backward citation search to mitigate this, we were surprised that our final corpus has a relatively small number of publications (17). Given our efforts, we believe the smaller number of publications is because this remains an emergent area of design and systems evaluation. We also hypothesize that there may be private or proprietary systems that are used by online platforms to identify and address online harassment. Because these are not described or evaluated in the research work, we have no insight into their design, victim-centeredness, or outcomes.

We chose to trade off a potential increase in corpus size that may have come from including AI prediction of toxicity and harassment for clarity in the analysis. In particular, our work focuses only on computationally-driven approaches *which interface with the victim as the user.* This focus excludes a large body of work on machine learning-based approaches to detect hateful or toxic content, a widespread strategy for computationally addressing harm[e.g., 5, 16, 27, 81, 108, 113]. As mentioned in Methods, we initially attempted to include these approaches to harassment in our analysis. However, the gaps in comparing setup, methods, and outcomes/evaluations created apples-to-oranges comparisons. Machine-learning based approaches are developed through training models on large datasets and evaluated by the accuracy of trained models, while system-based approaches are developed through design methods and evaluated by their effect on users.We acknowledge that improving algorithms that identify instances of online harassment can be a powerful approach to addressing online harassment [60].However, HCML researchers have pointed out that even machine learning approaches that are intended to be human-centered still risks dehumanizing victims' experience of harm and take away their agency and power [17]. One promising avenue of research that builds on both Jurgens et al. [60] and Chancellor et al. [17], as well as our work, would be to consider the extent to which machine-learning based approaches may adapt victim-centered principles in the design and evaluation of their models. Evaluating the intersection of victims-centered principles and these broader algorithmic strategies that do not directly involve victims as users (but can have material outcomes on them), and understanding how we might incorporate victims-centered principles into those strategies is a critical and exciting area of future work.

Finally, our focus on victims of online harassment as users while evaluating existing systems means that we do not systematically consider the role of other stakeholders, like moderators, harassers, and bystanders. Our findings, in fact, point to the value of including these stakeholders in design considerations–we saw systems where friends took on key roles as moderators, where victims wanted to see harassers informed and educated of their harms, and where surrounding community uplifted victims and helped them process the harassment they experienced. Future work to more comprehensively address online harassment will require understanding the roles and impacts of these stakeholders on how harassment is addressed. For example, evaluating whether and how do computational approaches take into account these various stakeholders can unearth gaps and opportunities for designing novel mechanisms.

6 CONCLUSION

We conducted a scoping literature review to evaluate how prior work on computational approaches to combat online harassment take into account identity traits of their users, their varying definitions of harassment and preferences for recourse after experiencing instances of harassment. Our analysis highlights the need to explicitly account for identity characteristics during the formative studies. We also observe a need for providing users with greater feedback when they report instances of harassment, and shifting or adding new mechanisms of classifications to better account for their subjective definitions of harassment. Finally, we highlight positive examples from our analysis on how systems leverage opportunities for collaboration and community-building to develop more meaningful ways to address harassment.

REFERENCES

[1] Emily a Vogels. 2021. The State of Online Harassment.

[2] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for Computing in Social Change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20).* Association for Computing Machinery, New York, NY, USA, 252–260. https://doi.org/10.1145/3351095.3372871

[3] Hilary Arksey and Lisa O'Malley. 2005. Scoping Studies: Towards a Methodological Framework. *International Journal of Social Research Methodology* 8, 1 (Feb. 2005), 19–32. https://doi.org/10.1080/1364557032000119616

[4] Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, San Jose California USA, 3895–3905. https://doi.org/10.1145/2858036.2858548

[5] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion.* ACM Press, Perth, Australia, 759–760. https://doi.org/10.1145/3041021.3054223

[6] Tanvi Bajpai, Drshika Asher, Anwesa Goswami, and Eshwar Chandrasekharan. 2022. Harmonizing the Cacophony with MIC: An Affordance-aware Framework for Platform Moderation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–22.

[7] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732.

[8] Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 conference on Designing interactive systems.* 93–102.

[9] Laura W. Black, Howard T. Welser, Dan Cosley, and Jocelyn M. DeGroot. 2011. Self-Governance through Group Discussion in Wikipedia Measuring Deliberation in Online Groups. *Small Group Research* 42, 5 (Oct. 2011), 595–634. https://doi.org/10.1177/1046496411406137

[10] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment Is Perceived as Justified. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (June 2018).

[11] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 24:1–24:19. https://doi.org/10.1145/3134659

[12] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–25. https://doi.org/10.1145/3359202

[13] Richard E. Boyatzis. 1998. *Transforming Qualitative Information: Thematic Analysis and Code Development.* Sage Publications, Inc, Thousand Oaks, CA, US. xvi, 184 pages.

[14] Johanna Brewer, Morgan Romine, and T. L. Taylor. 2020. Inclusion at Scale: Deploying a Community-Driven Moderation Intervention on Twitch. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20).* Association for Computing Machinery, New York, NY, USA, 757–769. https://doi.org/10.1145/3357236.3395514

[15] Jens Brunk, Jana Mattern, and Dennis M. Riehle. 2019. Effect of Transparency and Trust on Acceptance of Automatic Online Comment Moderation Systems. In *2019 IEEE 21st Conference on Business Informatics (CBI).* IEEE, Moscow, Russia, 429–435. https://doi.org/10.1109/CBI.2019.00056

[16] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. http://arxiv.org/abs/2010.12472 arXiv:2010.12472 [cs].

[17] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who Is the "Human" in Human-Centered Machine Learning: The Case of Predicting

Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–32. https://doi.org/10.1145/3359249

[18] Stevie Chancellor, Jessica Annette 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*. ACM, San Francisco California USA, 1201–1213. https://doi.org/10.1145/2818048.2819963

[19] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–30. https://doi.org/10.1145/3359276

[20] Kate Crawford and Tarleton Gillespie. 2016. What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society* 18, 3 (March 2016), 410–428. https://doi.org/10.1177/1461444814543163

[21] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 512–515.

[22] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25, 2 (April 2021), 700–732. https://doi.org/10.1007/s12119-020-09790-w

[23] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173986

[24] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D. Wherry, and Natalya N. Bazarova. 2018. Upstanding by Design: Bystander Intervention in Cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173785

[25] 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. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–26.

[26] Bryan Dosono and Bryan Semaan. 2020. Decolonizing Tactics as Collective Resilience: Identity Work of AAPI Communities on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–20. https://doi.org/10.1145/3392881

[27] Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. 2020. BERT and fastText Embeddings for Automatic Detection of Toxic Speech. In *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)*. IEEE, Tunis, Tunisia, 1–5. https://doi.org/10.1109/OCTA49274.2020.9151853

[28] Tina Ekhtiar, Armağan Karahanoğlu, Rúben Gouveia, and Geke Ludden. 2023. Goals for Goal Setting: A Scoping Review on Personal Informatics. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. ACM, Pittsburgh PA USA, 2625–2641. https://doi.org/10.1145/3563657.3596087

[29] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376293

[30] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (March 2006), 80–92. https://doi.org/10.1177/160940690600500107 Publisher: SAGE Publications Inc.

[31] Michelle Ferrier and Nisha Garud-Patkar. 2018. TrollBusters: Fighting Online Harassment of Women Journalists. In *Mediating Misogyny: Gender, Technology, and Harassment*, Jacqueline Ryan Vickery and Tracy Everbach (Eds.). Springer International Publishing, Cham, 311–332. https://doi.org/10.1007/978-3-319-72917-6_16

[32] Jerry Finn. 2004. A Survey of Online Harassment at a University Campus. *Journal of Interpersonal Violence* 19, 4 (April 2004), 468–483. https://doi.org/10.1177/0886260503262083

[33] Jesse Fox and Wai Yen Tang. 2017. Women's Experiences with General and Sexual Harassment in Online Video Games: Rumination, Organizational Responsiveness, Withdrawal, and Coping Strategies. *New Media & Society* 19, 8 (Aug. 2017), 1290–1307. https://doi.org/10.1177/1461444816635778

[34] Vicki S. Freimuth, Sandra Crouse Quinn, Stephen B. Thomas, Galen Cole, Eric Zook, and Ted Duncan. 2001. African Americans' views on research and the Tuskegee Syphilis study. *Social Science & Medicine* 52, 5 (2001), 797–808. https://doi.org/10.1016/S0277-9536(00)00178-7

[35] Seth Frey, P. M. Krafft, and Brian C. Keegan. 2019. "This Place Does What It Was Built for": Designing Digital Institutions for Participatory Change. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 32:1–32:31. https://doi.org/10.1145/3359134

[36] Batya Friedman, Peter H. Kahn, and Alan Borning. 2002. Value Sensitive Design: Theory and Methods. https://api.semanticscholar.org/CorpusID:18065345

[37] Doğa Gatos, Aslı Günay, Güncel Kırlangıç, Kemal Kuscu, and Asim Evren Yantac. 2021. How HCI bridges health and design in online health communities: a systematic review. In *Designing Interactive Systems Conference 2021*. 970–983.

[38] R. Stuart Geiger. 2016. Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space. *Information, Communication & Society* 19, 6 (June 2016), 787–803. https://doi.org/10.1080/1369118X.2016.1153700

[39] R. Stuart Geiger and Aaron Halfaker. 2013. When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes?. In *Proceedings of the 9th International Symposium on Open Collaboration (OpenSym '13)*. ACM, New York, NY, 6:1–6:6. https://doi.org/10.1145/2491055.2491061

[40] Tarleton Gillespie. 2018. "The Scale Is Just Unfathomable". https://logicmag.io/scale/the-scale-is-just-unfathomable/.

[41] Tarleton Gillespie. 2020. Content Moderation, AI, and the Question of Scale. *Big Data & Society* 7, 2 (July 2020), 2053951720943234. https://doi.org/10.1177/2053951720943234

[42] Nitesh Goyal, Leslie Park, and Lucy Vasserman. 2022. "You have to prove the threat is real": Understanding the needs of Female Journalists and Activists to Document and Report Online Harassment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3491102.3517517

[43] Maria J. Grant and Andrew Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal* 26, 2 (June 2009), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

[44] Kishonna Gray-Denson. 2012. Intersecting Oppressions and Online Communities. Information. *Information, Communication & Society* 15 (April 2012), 411–428. https://doi.org/10.1080/1369118X.2011.642401

[45] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17 (2015), 42–109.

[46] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–35. https://doi.org/10.1145/3479610

[47] Viviane Herdel, Lee J Yamin, and Jessica R Cauchard. 2022. Above and beyond: A scoping review of domains and applications for human-drone interaction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.

[48] Dorothy Howard and Lilly Irani. 2019. Ways of Knowing When Research Subjects Care. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300327

[49] Sohyeon Hwang and Aaron Shaw. 2022. Rules and Rule-Making in the Five Largest Wikipedias. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 347–357.

[50] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S. Ackerman, and Eric Gilbert. 2021. Yes: Affirmative Consent as a Theoretical Framework for Understanding and Imagining Social Platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–18. https://doi.org/10.1145/3411764.3445778

[51] Jane Im, Sarita Schoenebeck, Marilyn Iriarte, Gabriel Grill, Daricia Wilkinson, Amna Batool, Rahaf Alharbi, Audrey Funwie, Tergel Gankhuu, Eric Gilbert, and Mustafa Naseem. 2022. Women's Perspectives on Harm and Justice after Online Harassment. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 355 (nov 2022), 23 pages. https://doi.org/10.1145/3555775

[52] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376383

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

[54] Intel. 2022. Bleep Beta US - Intel Gaming Access. https://game.intel.com/giveaway/bleepbeta/

[55] 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 (Sept. 2019), 1–35. https://doi.org/10.1145/3338243

[56] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–21. https://doi.org/10.1145/3491102.3517505 arXiv:2202.08818 [cs]

[57] Shagun Jhaver, Seth Frey, and Amy X Zhang. 2023. Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms. *Social Media+ Society* 9, 4 (2023), 20563051231207857.

[58] 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 (April 2018), 1–33. https://doi.org/10.1145/3185593

[59] Shagun Jhaver, Alice Qian Zhang, Quanze Chen, Nikhila Natarajan, Ruotong Wang, and Amy Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. http://arxiv.org/abs/2305.10374 arXiv:2305.10374 [cs].

[60] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3658–3666. https://doi.org/10.18653/v1/P19-1357

[61] Os Keyes, Burren Peil, Rua M Williams, and Katta Spiel. 2020. Reimagining (women's) health: HCI, gender and essentialised embodiment. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–42.

[62] Charles Kiene and Benjamin Mako Hill. 2020. Who Uses Bots? A Statistical Analysis of Bot Usage in Moderation Teams. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–8. https://doi.org/10.1145/3334480.3382960

[63] Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–36.

[64] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 334:1–334:21. https://doi.org/10.1145/3476075

[65] Yubo Kou and Xinning Gui. 2020. Mediating Community-AI Interaction through Situated Explanation: The Case of AI-Led Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–27. https://doi.org/10.1145/3415173 arXiv:2008.08202

[66] 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*. ACM, Yokohama Japan, 1–12. https://doi.org/10.1145/3411764.3445279

[67] Robert E. Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press, Cambridge, MA.

[68] Travis Kriplean, Ivan Beschastnikh, and David W. McDonald. 2008. Articulations of Wikiwork: Uncovering Valued Work in Wikipedia through Barnstars. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA) *(CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 47–56. https://doi.org/10.1145/1460563.1460573

[69] Etienne G. Krug, James A. Mercy, Linda L. Dahlberg, and Anthony B. Zwi. 2002. The World Report on Violence and Health. *Lancet (London, England)* 360, 9339 (Oct. 2002), 1083–1088. https://doi.org/10.1016/S0140-6736(02)11133-0

[70] Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 548–559.

[71] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. Association for Computing Machinery, New York, NY, USA, 543–550. https://doi.org/10.1145/985692.985761

[72] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31, 2 (2014), 317–326. https://doi.org/10.1016/j.giq.2013.11.005

[73] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All That's Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 584–595.

[74] Calvin A Liang, Sean A Munson, and Julie A Kientz. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 2 (2021), 1–47.

[75] Alessandro Liberati, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P. A. Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. 2009. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine* 6, 7 (July 2009), e1000100. https://doi.org/10.1371/journal.pmed.1000100

[76] Amy Lyndon, Jennifer Bonds-Raacke, and Alyssa Cratty. 2011. College Students' Facebook Stalking of Ex-Partners. *Cyberpsychology, behavior and social networking* 14 (July 2011), 711–6. https://doi.org/10.1089/cyber.2010.0588

[77] Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. https://doi.org/10.1145/3173574.3174160

[78] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures. *New Media & Society* 19, 3 (March 2017), 329–346. https://doi.org/10.1177/1461444815608807

[79] Lavinia McLean and Mark D. Griffiths. 2019. Female Gamers' Experience of Online Harassment and Social Support in Online Gaming: A Qualitative Study. *International Journal of Mental Health and Addiction* 17, 4 (Aug. 2019), 970–994. https://doi.org/10.1007/s11469-018-9962-0

[80] Zachary Munn, Micah D. J. Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. Systematic Review or Scoping Review? Guidance for Authors When Choosing between a Systematic or Scoping Review Approach. *BMC Medical Research Methodology* 18, 1 (Nov. 2018), 143. https://doi.org/10.1186/s12874-018-0611-x

[81] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Montréal Québec Canada, 145–153. https://doi.org/10.1145/2872427.2883062

[82] Giovanna Nunes Vilaza, Kevin Doherty, Darragh McCashin, David Coyle, Jakob Bardram, and Marguerite Barry. 2022. A Scoping Review of Ethics Across SIGCHI. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 137–154. https://doi.org/10.1145/3532106.3533511

[83] Giovanna Nunes Vilaza, Kevin Doherty, Darragh McCashin, David Coyle, Jakob Bardram, and Marguerite Barry. 2022. A scoping review of ethics across SIGCHI. In *Designing Interactive Systems Conference*. 137–154.

[84] Diana Papaioannou, Anthea Sutton, Christopher Carroll, Andrew Booth, and Ruth Wong. 2010. Literature searching for social science systematic reviews: consideration of a range of search techniques. *Health Information & Libraries Journal* 27, 2 (2010), 114–122. https://doi.org/10.1111/j.1471-1842.2009.00863.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1471-1842.2009.00863.x.

[85] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 52–63. https://doi.org/10.1145/3442188.3445870

[86] Laura Parson. 2019. *Considering Positionality: The Ethics of Conducting Research with Marginalized Groups*. Springer International Publishing, Cham, 15–32. https://doi.org/10.1007/978-3-030-05900-2_2

[87] Elissa M. Redmiles, Jessica Bodford, and Lindsay Blackwell. 2019. "I Just Want to Feel Safe": A Diary Study of Safety Perceptions on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* 13 (July 2019), 405–416.

[88] Elizabeth Reid, Regan L. Mandryk, Nicole A. Beres, Madison Klarkowski, and Julian Frommel. 2022. Feeling Good and In Control: In-game Tools to Support Targets of Toxicity. *Proceedings of the ACM on Human-Computer Interaction* 6, CHI PLAY (Oct. 2022), 235:1–235:27. https://doi.org/10.1145/3549498

[89] Sarah T. Roberts. 2014. *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation*. Ph. D. Dissertation. University of Illinois at Urbana-Champaign, United States – Illinois.

[90] Darcell P Scharff, Katherine J Mathews, Pamela Jackson, Jonathan Hoffsuemmer, Emeobong Martin, and Dorothy Edwards. 2010. More than Tuskegee: understanding mistrust about research participation. *Journal of health care for the poor and underserved* 21, 3 (2010), 879.

[91] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 155:1–155:27. https://doi.org/10.1145/3274424

[92] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from Justice Theories to Support Targets of Online Harassment. *New Media & Society* 23, 5 (May 2021), 1278–1300. https://doi.org/10.1177/1461444820913122

[93] Nick Seaver. 2021. Care and Scale: Decorrelative Ethics in Algorithmic Recommendation. *Cultural Anthropology* 36, 3 (Aug. 2021), 509–537–509–537. https://doi.org/10.14506/ca36.3.11

[94] 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*. ACM, Portland Oregon USA, 111–125. https://doi.org/10.1145/2998181.2998277

[95] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society* 21, 7 (July 2019), 1417–1443. https://doi.org/10.1177/1461444818821316

[96] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph 'Jofish' Kaye. 2005. Reflective Design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility (CC '05)*. Association for Computing

Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/1094562.1094569

[97] Helen Sharp, Jenny Preece, and Yvonne Rogers. 2019. Interaction design: beyond human-computer interaction.

[98] Tamara Shepherd, Alison Harvey, Tim Jordan, Sam Srauy, and Kate Miltner. 2015. Histories of Hating. *Social Media + Society* 1, 2 (July 2015), 2056305115603997. https://doi.org/10.1177/2056305115603997

[99] Ellen Simpson and Bryan Semaan. 2021. For You, or For"You"? Everyday LGBTQ+ Encounters with TikTok. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (Jan. 2021), 252:1–252:34. https://doi.org/10.1145/3432951

[100] Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Mediating Conflicts in Minecraft: Empowering Learning in Online Multiplayer Games. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 2696–2707. https://doi.org/10.1145/3025453.3025516

[101] Jean Y. Song, Sangwook Lee, Jisoo Lee, Mina Kim, and Juho Kim. 2023. ModSandbox: Facilitating Online Community Moderation Through Error Prediction and Improvement of Automated Rules. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3544548.3581057

[102] Evropi Stefanidi, Marit Bentvelzen, Paweł W Woźniak, Thomas Kosch, Mikołaj P Woźniak, Thomas Mildner, Stefan Schneegass, Heiko Müller, and Jasmin Niess. 2023. Literature Reviews in HCI: A Review of Reviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–24.

[103] Francesca Stevens, Jason R. C. Nurse, and Budi Arief. 2021. Cyber Stalking, Cyber Harassment, and Adult Mental Health: A Systematic Review. *Cyberpsychology, Behavior and Social Networking* 24, 6 (June 2021), 367–376. https://doi.org/10.1089/cyber.2020.0253

[104] Sharifa Sultana, Mitrasree Deb, Ananya Bhattacharjee, Shaid Hasan, S.M.Raihanul Alam, Trishna Chakraborty, Prianka Roy, Samira Fairuz Ahmed, Aparna Moitra, M Ashraful Amin, A.K.M. Najmul Islam, and Syed Ishtiaque Ahmed. 2021. 'Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–18. https://doi.org/10.1145/3411764.3445154

[105] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 247–267. https://doi.org/10.1109/SP40001.2021.00028

[106] Andrea C. Tricco, Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D.J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A. Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G. Wilson, Chantelle Garritty, Simon Lewin, Christina M. Godfrey, Marilyn T. Macdonald, Etienne V. Langlois, Karla Soares-Weiser, Jo Moriarty, Tammy Clifford, Özge Tunçalp, and Sharon E. Straus. 2018. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine* 169, 7 (Oct. 2018), 467–473. https://doi.org/10.7326/M18-0850

[107] Anna Lowenhaupt Tsing. 2012. On Nonscalability: The Living World Is Not Amenable to Precision-Nested Scales. *Common Knowledge* 18, 3 (Aug. 2012), 505–524. https://doi.org/10.1215/0961754X-1630424

[108] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. http://arxiv.org/abs/1809.07572 arXiv:1809.07572 [cs].

[109] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1231–1245. https://doi.org/10.1145/2998181.2998337

[110] Nina Wallerstein, Michael Muhammad, Shannon Sanchez-Youngman, Patricia Rodriguez Espinosa, Magdalena Avila, Elizabeth A. Baker, Steven Barnett, Lorenda Belone, Maxine Golub, Julie Lucero, Ihsan Mahdi, Emma Noyes, Tung Nguyen, Yvette Roubideaux, Robin Sigo, and Bonnie Duran. 2019. Power Dynamics in Community-Based Participatory Research: A Multiple–Case Study Analysis of Partnering Contexts, Histories, and Practices. *Health Education & Behavior* 46, 1_suppl (2019), 19S–32S. https://doi.org/10.1177/1090198119852998 arXiv:https://doi.org/10.1177/1090198119852998 PMID: 31549557.

[111] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, 138–142. https://doi.org/10.18653/v1/W16-5618

[112] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[113] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a Lexicon of Abusive Words – a Feature-Based Approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1046–1056. https://doi.org/10.18653/v1/N18-1095

[114] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–26. https://doi.org/10.1145/3449280

[115] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. https://doi.org/10.1145/3491102.3517614