Taking the control back – An adventure in developing personalized content moderation

Wenshan Luo*
Computer Science and Engineering
University of California San Diego
San Diego, California, USA
w1luo@ucsd.edu

Pat William Pannuto
Computer Science and Engineering
University of California San Diego
La Jolla, California, USA
ppannuto@ucsd.edu

Kristen Vaccaro
Computer Science and Engineering
University of California San Diego
San Diego, California, USA
kvaccaro@ucsd.edu

Abstract

Online platforms are riddled with harassment, which significantly impacts the well-being of users. Unfortunately, the content moderation solutions provided by platforms often disappoint end-users as they fail to equip individuals with sufficient controls for their personal situations. In this work, the author, who personally experienced a sustained harassment campaign on Twitter, decided to regain control by constructing an automated, personalized, and collaborative anti-harassment system to protect herself, which has proven itself to be effective. The experience of developing—and re-developing in the face of repeated platform API changes and restrictions—this personalized content moderation system highlights many design issues that make managing severe online harassment such a challenge and invites critical study of the power dynamics between large online platforms and individual users. Through this analysis, this report aims to inform better designs to help platforms more effectively protect victims.

CCS Concepts

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

Keywords

online harassment; personalized content moderation; anti-harassment tool; autoethnography

ACM Reference Format:
Wenshan Luo, Pat William Pannuto, and Kristen Vaccaro. 2026. Taking the control back – An adventure in developing personalized content moderation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3772318.3791905

1 Introduction

Social media platforms are consistently plagued by online harassment [80, 98, 140], which significantly impacts the user experience and the well-being of victims [51, 93, 120, 126]. The problem poses a significant challenge to online platform governance [14, 36, 100], and users often find the content moderation solutions adopted by platforms to be unsatisfactory [52, 84, 112]. Because of frustrations with the platforms, anti-harassment tools that give the community and the users more control have emerged in recent years [13, 15, 16, 37, 86, 92], and users have expressed interest in personalizing their content moderation experience more [70, 72].

This work explores an extreme case of personalized content moderation, developed in response to a prolonged harassment campaign.This work adopts an autoethnographic approach, similar to other ethnographic work on online harassment [71, 114, 135]. In this case, the target of the harassment campaign and the designer of the anti-harassment system happen to be the same person—me.

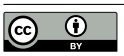
It draws on the primary author's personal experience of enduring stalking and targeted harassment on Twitter[1] for over nine months. A single harasser created numerous accounts to target me and my friends with up to dozens of attacks each day, taking advantage of unintended uses of platform tools to increase my exposure to the abuse. The harassment campaign took place December 2022 to September 2023, coinciding with a period that saw a dramatic increase in hate speech in general on the platform [58]. This work will present observations on the context and experience of a targeted harassment campaign, drawing together autoethnographic reflections as well as quantitative data.

In response to the harassment campaign, I constructed a personalized automated anti-harassment system as a defense. The system filters all incoming interactions directed at me by employing an account-based approach, which is especially effective when the harasser creates hundreds of accounts with identifiable behavioral patterns. This system minimizes my exposure to the content from the harasser and reduces the effort of reporting significantly. Additionally, it enables collective action, inviting allies to join in the fight against the harasser by automating their accounts for blocking and reporting as well. The system has effectively protected me during the campaign, resulting in the removal of over 96% of the harasser's accounts thus far. The comprehensive automated logging enabled by this protection system also serves as the primary documentation used for the reflection presented in this work.

Development of the anti-harassment system continued over a period of dramatic changes to the Twitter API, which used to be the primary means through which developers made customized tools for the platform. This work will discuss the challenges that arose from these unanticipated and restrictive platform changes and the requisite shifts in approach and capability of the anti-harassment

*"I" in this work refers to this person, who experienced the targeted harassment campaign and developed the protection tools; the other authors assisted with the analysis and preparation of this manuscript.

[1]Twitter was renamed X on July 23, 2023. However, we retain the original name for clarity and because most of the author's use of the platform occurred under its original name.



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791905

tool. The paper concludes with recommendations for anti-abuse designs drawn from these experiences.

The contributions of this work include: (1) a detailed account of the personal and interpersonal challenges of managing harassment on Twitter, including the cognitive and emotional work involved, (2) empirical insights from the autoethnography study, (3) design goals of a personalized anti-harassment systems, (4) a research prototype of a personalized anti-harassment system and a discussion of its accessibility, (5) reflections on the benefits and challenges of being a victim-developer, (6) a critique on designing important anti-harassment tools for social media platforms in a post-API era, and (7) ethical discussion for conducting autoethnographic research on sensitive and emotionally demanding topics.[2]

2 Background

Online harassment is a common experience of internet users across platforms and across national borders [19, 80, 83, 98, 100, 102, 144]. In the US, a 2021 Pew Research Center survey found that 41% of Americans have personally experienced online harassment and 25% have experienced severe forms of harassment which include physical threats, stalking, sexual harassment, and sustained harassment [140]. Online harassment takes a heavy toll on its targets [126], and women and people from marginalized communities are especially vulnerable [54, 84, 108, 139, 140].

A common strategy for dealing with online harassment is content moderation, which has been defined as "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [53]. Most platforms have content moderation policies to address online harassment, which often differ significantly from one platform to another [100]. Content moderation is critical to the operation of online platforms, so much so that Gillespie argues that moderation is actually the main commodity offered by a platform [45].

There are many approaches to content moderation, a complete summary of which is beyond the scope of this work. The trade-offs of different approaches have been thoroughly reviewed recently by Jiang et. al [73]. The rest of this section provides an overview of the decisions and insights most germane to this work.

2.1 Experiences with Online Harassment

The severity of online harassment ranges from one-off comments to prolonged and repeated abuse [80, 125]. Regardless of severity, targets of online harassment experience emotional and physical distress, with disruptions in both their online and offline life [13, 126]. While only a minority of harassment cases involve severe, sustained harassment, such cases can inflict significant psychological and social harm on those targeted. In 2006, Megan Meier's tragic death after severe online harassment and bullying shocked the world and highlighted the link between severe harassment and suicide [23, 59]. Despite the many years since, severe harassment remains a problem. A recent report detailed a prolonged, cross-platform campaign targeting Asian American academics with thousands of insults, false title IX complaints, and death threats, which left many targets fearing that online abuse could escalate into physical harm [57].

[2]An anonymous reviewer initially articulated this list of contributions, which we reproduce here — in lightly edited form — with appreciation.

Targets of harassment often rely on platform-provided tools such as content removal, banning, and labeling. Researchers have also studied additional approaches, such as requiring apologies, rating systems, public shaming, and even payments to manage harassment [65, 143]. However, the platform-provided tools can be ineffective and frustrating to use [13, 52]. As a result, many targets of harassment turn to other strategies for mitigation, include self-censorship, impression management, and disengagement (e.g., account deletion) [24, 55, 126, 139]. In some cases, targets have been able to leverage third-party bots and community action (e.g., virtual walkouts) to deter harassment [56]. The harassment campaign described in this work is on the severe end of the spectrum; according to Thomas et al.'s taxonomy of online harassment [125], my experience involved five of the seven major categories of online harassment: toxic content (e.g., bullying, trolling), content leakage (e.g., doxxing), overloading (e.g., brigading), surveillance (e.g., stalking), and impersonation, with only false reporting and lockout excluded.

2.2 Methods for Content Moderation

Moderation decisions can be automated, manual, or a combination [82]. My system draws from both threads of work.

2.2.1 Automated Content Moderation. Automated content moderation relies heavily on machine learning based techniques [34, 76, 91], which though efficient have performance issues and biases [29, 50, 62]. While the adoption of automated decision-making by platforms as a scalable solution for content moderation raises numerous questions [46, 85], most research has focused on the experience of moderators [111, 116]. Much less has focused on the legitimate use of automated tools by *users* to alleviate the burden of moderation.

Individuals experiencing severe online harassment often find themselves with no alternative but to spend hours on blocking and reporting. As a result, many simply give up [13]. Gillespie, while generally critical of automated content moderation, argues that the human costs of reviewing harmful content might be the strongest justification for automation [46]. User-developed bots have been used for content moderation on Wikipedia [44], Reddit [69], and Twitch [56]. Geiger theorizes that this phenomenon challenges the very concept of a platform, because there is no longer a stable, monolithic "platform" when user code runs on independent machines and plays a significant role in system governance. While bots and decentralized code bring benefit to communities, they also introduce unique challenges for governance and sustainable development [42]. This work aims to reify these tensions by critically examining the experience of one individual who—amidst an active harassment campaign—had to uncover both the technical and policy reasons that platform-provided automated moderation was insufficient (Section 4) and who then had to (repeatedly) design and develop their own automated moderation system that reached beyond the nominal programmable surface provided by the platform (Section 5).

2.2.2 Manual Content Moderation. An alterative approach to automation attracting growing interest from researchers is community-based or collaborative content moderation. For example, in the form

of community-curated blocklists, which were widely used when the Twitter API supported them [43, 71]. Researchers have also built tools to mobilize friends of a harassment target, or volunteers, to collectively combat online harassment and provide support to the target [13, 37, 86]. Communities have also self-organized, for example, battling harassment campaigns by coordinating virtual walkouts [56]. In this work, I mobilize my allies not only to support me but also to mitigate the social harm caused by the harassment — when my friends help me, the harassment loses much of its power for social damage.

Traditionally, however, the primary mechanism for manual content moderation was reporting. The process begins when a user flags problematic content and submits a report for moderators to review. Reporting imposes non-trivial cognitive and emotional load on the user, and it is often slow and frustrating [13, 52], as I repeatedly experienced. Some platforms have established trusted reporter partnerships to prioritize certain reports [49]. However, even these partnerships are rarely effective at removing content [89].

Some have argued that platforms deliberately employ thin and ambiguous flags in the reporting process to avoid responsibility and suppress dissent [25].

While reporting can serve as a useful tool in combating online harassment, it can also be misused as a means of repression or serve other instrumental purposes [38, 79, 145]. Therefore, when making the library that empowers this work publicly available, I have omitted documentation for reporting. To minimize potential abuses of reporting systems, Crawford and Gillespie propose an open backstage model that preserves the history of debates pertaining to specific problematic content [25]. While this approach could be a suitable alternative for news discussions, it may not be appropriate for dealing with defamatory remarks and highly personal insults targeted at individuals. Indeed, for me, a central design goal was for the harassing content to never be seen by me or my followers (Section 5).

2.3 Personalized Content Moderation

Personalized content moderation — where the user can configure the content moderation according to their own preferences [70] — is of widespread interest among the public [72]. Limited forms of personalization are integrated into some platforms, like custom word lists to filter comments [90] or the ability to mute certain notifications [130]. Twitter also briefly promoted paid third party content moderation tools such as BlockParty [15] (no longer available), Moderate [92] (no longer available), and Bodyguard [16] (intended for brands); these allowed for more end-user customization in 2022, but no longer. The experience examined in this work provides a uniquely holistic view on personalized moderation, as it first (Section 4) considers an individual's interactions, experiences, and power through the limiting scope of the end-user-immutable personalized moderation tools afforded by a platform, and then (Section 5) opens the door to mutable personalized moderation tools and explores the lived experience of an iterative design cycle across a greenfield consideration of desired moderation capabilities, the actual capabilities exposed by the platform, and the technical capacity to develop new capabilities along with the stimuli—both highly individual (i.e., the harasser changing behavior in response to new moderation capabilities) and systemic (i.e., the changing or removal of platform features)—that propel this cycle.

2.4 Major Moderation Actions Available to Twitter End-users

Twitter implements a standard array of moderation actions for end-users. Here we focus on their functionality; the limitations of these actions are discussed in Section 6.

The following example illustrates how these actions function: *Someone begins harassing Alex on Twitter. The harasser leaves an abusive comment on Alex's post. There is no way for Alex to delete the comment. But Alex could mute or block the harasser.*

If Alex **mutes** the harasser, the harasser will not know about it. The harasser can still read Alex's posts and interact with them, but Alex will not be notified about the interactions. When Alex reads the post with the harasser's comment, the comment will not be shown. Instead, a banner will show "*This post is from an account you muted*," and clicking on the banner will reveal harasser's comment.

If Alex **blocks** the harasser, the harasser will soon know. If the block occurs as the harasser posts another comment, the submission will fail and the harasser will see a message, "*The post you are trying to reply to has been deleted or is not visible to you*". If the harasser visits Alex's homepage, all posts are hidden and a large "*You're blocked*" message is shown instead. In theory, a block should prevent all future interactions from the harasser, but, in practice, existing bugs[3] allow the harasser to quote Alex and participate in the discussion under Alex's posts. Even though the harasser's comment is now invisible to Alex, it is still visible to everyone else.

If the harasser registers a new account and continues the harassment, Alex may choose to make their own account **protected**. In that case, the harasser can only see or interact with Alex's posts if they follow Alex, which requires Alex's approval. However, the visibility restriction does not apply only to the harasser. With a protected account, Alex can no longer engage with non-followers, since any engagements are only visible to Alex and Alex's followers.

While there are other ancillary features (e.g., mutes can also be used to remove particular words, usernames, hashtags, etc.), muting, blocking, and protecting accounts are the primary capabilities for users to moderate what they see. These capabilities are not static; on August 18, 2023, Twitter announced that the blocking function might be removed, which provoked a backlash from the users [127]. Blocking has not been removed at the time of submission. Appendix A provides a comparison to other popular platforms.

3 Autoethnographic Approach

This work employs an autoethnographic approach, which positions the author's subjective experience and self-observation at the center of the investigation [105]. Complex social issues are difficult to grasp solely through quantitative means, without considering how they unfold in the real world and are lived in practice [1]. Online harassment exemplifies this challenge: platform-level metrics and

[3]Once a harasser has a target tweet's URL (which they can use another account, or simply log out, to obtain), they can post that URL directly in the new tweet form to create a quote tweet on Twitter, even after being blocked. Similarly, the harasser can use another account to obtain the URL of any replies to the target's tweet. The harasser can directly visit the URL and talk to people replying to the original tweet, even while still using the blocked account.

aggregate statistics cannot capture the lived experience of sustained hostility and its impact on individuals. Autoethnography, by centering first-hand accounts, supplies the context and interpretive depth needed to understand the dynamics involving the harasser, the platform, the victim, and the developer (same as the victim in this work). The approach taken here aligns more closely with analytical autoethnography, as this work seeks to use empirical data to gain insight into the broader phenomenon of harassment [117].

Ideally this approach can "share voices that might not otherwise have been heard, and presents insights that might otherwise have been too subtle to elicit" [26]. However, there are also criticisms about its objectivity and reliability [105]. Recently, autoethnography has been employed in HCI work, such as studies of long-term self-tracking [61], cybersecurity practices [129], and even pursuing posthuman design through bird watching [10].

In the context of content moderation, autoethnography has been used in several previous studies. Many earlier researchers relied heavily on manual notes for their autoethnographic work. For instance, Carolina Are documented her experiences of being shadowbanned for nudity on major social media platforms using manual notes [6, 7]. In contrast, my automated harassment prevention tool generates documentation and notes in great detail at scale.

Some researchers have focused on studying the harassment experiences of specific groups. For example, a researcher at The Guardian offers an insider's perspective on the online hostility faced by journalists within comment sections, which provides insights into how these journalists collectively experience and perceive abusive comments [41]. Other researchers have also shared first-person accounts of coordinated harassment campaigns, like Gamergate, detailing the harassment encountered while researching and discussing the issue [141]. Another described the experience of receiving Twitter trolling from a group of accounts mocking academics [20]. My experience is different, as my interactions have led me to believe that I am dealing with a single, obsessive stalker.

3.1 Positionality Statement

I am an Asian woman with a technical background, completing a graduate degree in Computer Science, who developed my own personalized content moderation system to deal with a sustained harassment campaign on Twitter. I currently reside outside my home country, where Twitter is blocked by the government and has no in-country operation.[4]

My coauthors are faculty in my department, including my advisor, and neither has personally experienced a harassment campaign like the one I describe. After I expressed in June 2023 that I wanted to share my experience in a meaningful way, my advisor encouraged me to develop a manuscript that could contribute my insights to the broader research community and connected me, given that my primary research area is systems, to a coauthor who specializes in social computing. This coauthor provided guidance on how my experience fits into, reflects on, and extends prior work. Both coauthors supported me as I worked through this traumatic experience.

This work differs from previous studies in several key aspects. Note-based efforts provide limited quantitative data on harassment experiences, while the automated logging from the anti-harassment system I developed and deployed enables much more extensive documentation of harassment campaigns and anti-harassment actions than was previously possible; critically, this automation creates authoritative, time-accurate records for events—including those only found to be significant on retrospective analysis—that I am able to use to support the reflection and analysis throughout this study. Additionally, unlike prior work, this study uniquely focuses on a targeted harassment campaign by what I believe to be a single stalker for unknown reasons. To supplement the autoethnographic reporting, this work also employs systematic experimentation with different platforms' content moderation tools to uncover subtle, yet surprising, aspects of their operation that impacts user experience.

As a victim-developer, several things make my observations and reflections unique. Even though I have been a social media user for more than a decade, being a harassment victim has transformed my experience of social media: the heightened sense of danger enables me to perceive nuances of platform design that I would not have noticed otherwise. I now constantly ask myself while using any social network site: can this functionality be exploited by a harasser? What counter-measures can I use if my harasser follows me here? Checking privacy and safety-related settings has become a second nature when exploring new platforms.

As a tech-savvy user, my background knowledge of modern web technologies helps me understand Twitter and other social media platforms better. Being the developer of my own self-defense system, my experience directly informs the design and evolution of the system. I am able to respond to any issues of the system without delay, dealing with all the practicalities of system development and maintenance myself. I was also able to tailor the data collection process to minimize any possible trauma. This level of timeliness and control is unavailable to most general users today.

Prior to the events of this harassment campaign, I was well aware of the general online harassment problem. I had voluntarily reported state-sponsored trolls harassing journalists on Twitter in the past. However, this is the first time that I have experienced sustained and severe online harassment myself. I seek to empower targets of severe online harassment through sharing my experience. To that end, I have also made the library at the core of the anti-harassment system available upon request.

Throughout this work, I take a critical approach towards interpreting the reported experience, recognizing that meaning is shaped not only by my individual perspective but also by structural conditions and power relations. My analysis is informed by intellectual traditions such as platform studies, critical algorithm studies, and related strands of scholarship that foreground the political economy and power dynamics of sociotechnical systems, highlighting how personal experience is shaped not only by technological design but also by the economic imperatives and governance structures that underpin platform operations.

The decision to publish did not alter the design of the system; it only required that I preserve all existing logs for analysis.

3.2 Method

This work, which covers a period of harassment and anti-harassment efforts by the first author ("I") spanning from December 28, 2022

[4]This makes dealing with the harasser through legal means impractical and means I focus more on improving my experience than preserving evidence.

to October 1, 2023 on Twitter, was not pre-planned and was not intended as a research project; it was later, when encouraged to share my story, that I began to reflect on my experience in a more systematic way.

In line with established practices in autoethnography, I document my experiences reflexively through system-generated artifacts, recalling, and self-observation, and I treat my own perspective as an interpretive instrument.

As I believe that I am dealing with an obsessive, stalking harasser who has never had any benign interactions with me, in this work, I treat *all* interactions (including follows, comments, likes, reposts, and quotes) from the harasser's accounts as harassment. While the harasser's comments, reposts and quotes are stored temporarily during my system's processing pipeline, I do not permanently store—and will not release—any of this content for privacy reasons.

Rather than relying on a manual journal, this work primarily uses the logs generated automatically by my anti-harassment system, which documents metadata for every Twitter interaction (including follows, comments, likes, reposts, and quotes) I received. The recorded metadata includes the timestamp, interaction type, and user metadata, such as numerical user ID, user creation timestamp, following and followers count, tweet count, and favorite count at the time of interaction. For debugging and system monitoring purposes, the activity logs of the system, which record when the bots were running and what actions were taken, are also saved. I have deliberately chosen not to record the content of the interactions to avoid reading the harasser's abusive messages when reviewing the logs (see section 6.5 and section 5.1). The system also documents every report made by my friends' and my accounts, including the IDs of the reported tweets and accounts, the types of reports (currently: `Spam`, `Targeted Harassment`, `Insulting`, `Inciting Harassment`, `Threatening to Expose`, `Violent Speech`, `Wish of Harm`, `Threatening with Violence`), and tracks the status changes (normal, restricted, or removed) of the reported accounts over time. The logs allow me to assess both the harasser's behavior and how effectively my system protected me.

Out of a sense of urgency, the early anti-harassment system was temporarily deployed on a server hosted by my department, as finding a suitable free-tier server for hosting took some time. Eventually, I was fortunate to obtain a powerful server from a major cloud provider at no cost as part of their campaign to attract developers (a server with similar specifications would cost more than $40 per month on AWS). While my system could run on typical free-tier virtual machines with only 1 GB of RAM, this powerful free server made experimenting easier, like using my system to monitor and report harassers in other harassment campaigns, replicating my system for friends who need help, and running local natural language processing tasks such as trying out self-hosted large language models. After January 21, 2023, I migrated the system to the new server. The original logs before the migration were not preserved, therefore screenshots of the logs I made during the period (see Fig. 1 as an example) are used to fill in the gap. I had captured the screenshots to demonstrate the extent of the harassment and to alert my friends.

The logs are intact from January 21, 2023 onwards, with June 12–13, 2023 being exceptions. On June 12, Twitter disabled my access to the account activity API, and I had to transition fully to client notification-based logging (see Section 5.4.1 for a discussion), which was operational by June 14.

Where relevant, text conversations with friends and my own tweets about the harassment campaign are also used for recollection and reflection. Finally, code change logs and commit messages from software version control (git) of my anti-harassment tool are used to recall its major changes and to add occasional additional details.

Drawing on the logs and recollections detailed above, this paper takes a narrative approach, weaving together documented experiences with analysis and reflection, and focuses on my personal experiences with the harassment campaign as well as the development and maintenance of the anti-harassment system. Descriptive statistics of the harassment campaign and the anti-harassment effort are provided to support these narratives. Based on these narratives, I take an inductive approach to arrive at the insights about platform design.

As this work is based on my personal experiences, it has inherent limitations. To solicit feedback on the findings, I followed the common practice of member checking [12] by sharing a two-page overview of the study's methodology and results with academics who had publicly described their experiences of online harassment; my university's IRB determined that this did not require review. I received 5 email responses. Their insightful feedback helped me refine the discussion and strengthen the interpretation of the results. The specifics of my own situation are acknowledged in the positionality statement. As a tech-savvy user, my experience may not represent those of users with a non-technical background. However, the fact that even as a tech-savvy user, equipped with considerable resources, I struggled to address a (presumed) single stalker highlights the challenges of combating targeted online harassment.

3.3 Ethical Considerations and Further Reflections

While I have taken reasonable precautions, as explained below and throughout this document, I am nonetheless aware of and have considered the potential risks associated with publishing this work. Specific risks include: (1) becoming an open target for online harassment communities, (2) retaliation by the harasser(s), (3) the possibility of my tight-knit community being identified and infiltrated, (4) authorities in my home country using this work as evidence that I operate a Twitter account, an act that is technically illegal, though rarely prosecuted, and (5) Twitter potentially banning my account and taking legal action against me for violating its Terms of Service (ToS). I have made up my mind to publish the work despite the risks.

Still, I do aim to minimize any risks. To minimize the privacy risks, this paper does not disclose any explicit information about the harasser(s) or my friends' online or offline identities. It also excludes any tweets or exchanges between me and the harasser(s) or my friends. As a result, the likelihood that someone without access to Twitter's internal data could identify my or my friends' accounts is low. However, it is still possible for someone working at Twitter to identify me, based on disclosures such as my unusually large number of submitted reports and their timing.

The incident was emotionally difficult, and managing the ongoing harassment took a psychological toll. At the same time, building

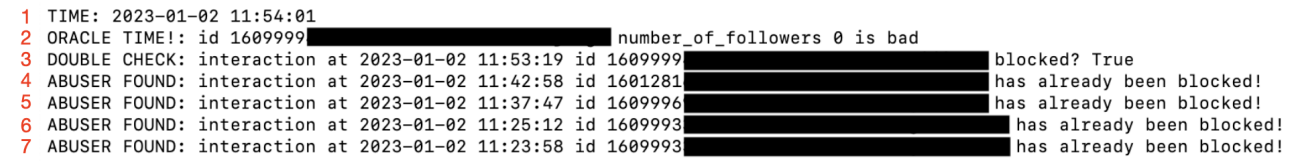


Figure 1: Redacted screenshot (revealing only the first 7 digits of the numerical Twitter IDs) of the system operation log that I shared with my friends on January 2, 2023. At that time, my detection bot was running every two minutes. The screenshot displays the output of the bot running on January 2 at 11:54:01 (line 1). During this run, a new harassing account was identified (line 2) and blocked (line 3) within one minute of its interaction with me. Logs also track interactions from three other recently blocked accounts (line 4-7).

the system, analyzing the logs, and documenting the experience helped me regain a sense of agency. Writing the manuscript was sometimes challenging, particularly since this work falls outside conventional publication norms. Throughout this period, I relied on a small support network that included friends on Twitter and my coauthors. Their encouragement and willingness to discuss both the emotional and technical aspects of the experience were important forms of support.

In preparing this manuscript, I consulted with my university's IRB, which concluded that this work, including the member checking component, did not require IRB review. I notified all of my Twitter followers, including the friends who assisted me, all of whom also use anonymous accounts, that I am sharing my experience in the form of a paper. I shared parts of the drafts as well as analysis in the form of plots and diagrams with my followers at several stages; they were shocked by the full scale of the harassment campaign and the tenaciousness of the harasser(s). None expressed concerns with publishing.

The trade-offs I made when designing and using the system are deeply personal. I prioritized minimizing my exposure to anything from the harasser(s), and I have no regrets about missing their content. However, other individuals may make different choices. Some may fear missing out on potential threats, while others may be concerned about isolating themselves [52]. Fortunately, my protection system can be customized to accommodate these needs.

From Twitter's standpoint, my reverse engineering, as well as account automation, could be seen as a violation of the ToS and as a form of abuse itself. My friends and I risk having our Twitter accounts banned for ToS violation; I explained this before deploying my system, and my friends agreed to help me despite the risk. However, I believe that when the platform fails to ensure the safety of its revenue-generating users by failing to enforce its own policies, the user has the right to take measures for self-protection, so long as it does not harm other legitimate users. Indeed, it has been argued that, when benefits to society outweigh the harm to the company, researchers should be allowed to violate terms of service [134], and the United States federal court has ruled that such violations are not crimes in Sandvig v. Barr [33], though given the complex international regulatory landscape, this protection only extends to researchers based in the US. As I develop a free tool to combat harassment—harassment that is pervasive on Twitter, where the platform's existing tools fall short of meeting users' needs [89]—I argue that the benefits far outweigh any potential risks. Since the discontinuation of the old Twitter API, users are left with only the official tools unless they are willing to pay the steep price for the now-unaffordable API. My tool offers a robust alternative with enhanced functionality. However, its release is not without risks. For instance, the asynchronous reporting feature, which allows parallelized reporting, could be exploited for malicious reporting campaigns. To minimize the risk of misuse, I have designed the library so that potentially harmful code is generated through manipulation of the abstract syntax tree, making it inaccessible via standard code inspection. I do not share the details or usage of the reporting functionality without first verifying the requester's good intentions. At the same time, anyone capable of discovering how to invoke these hidden functions on their own would already have the technical expertise to reverse-engineer the same endpoints independently, and it would likely be faster for them to do so directly than to analyze my code.

When I share my library online, there is a possibility that loopholes will be found by the harasser(s) to circumvent the protections. There has long been debate within the open source community about the security implications of open source [101]. In my situation, since the specific configuration of the system is kept locally, harassers only have access to the system's general logic. Reverse-engineering parameters would require extensive experimentation by harassers.

The system, which has been extensively customized to meet my specific needs, has proven largely effective in mitigating attacks from my harasser(s). However, it does not address the underlying issue of the ongoing harassment campaign, as it does not render it impossible for the harasser(s) to target me on Twitter. I believe that if Twitter were better designed, I might not have to rely on (nor repeatedly re-engineer) this system at all.

4 The Harassment Campaign

In responding to the harassment campaign, I inevitably use my observations to form hypotheses about the harasser(s)'s behavioral patterns and motivations. Many of my observations are supported by quantitative data, while others are best-effort guesses from a victim's perspective. I use words like "seems," "likely," or "may" to indicate observations of subjective nature. As these suppositions shape and inform my experience of the harassment, how I respond to my harasser, and my approach to building the system, I include them in this work as an essential part of the autoethnography.

As one specific example, based on highly consistent harassing behavior, I strongly believe that the entire harassment campaign

was carried out by a single individual. Subsequent text will omit the qualifiers (i.e., 'my harasser' not 'my harasser(s)') around the uncertainty of this fact, as my mental state throughout the experience reflected that of combating a single individual.

I have no concrete proof of the harasser's identity, but I suspect they may be a woman from my home country with whom I briefly interacted on a different social media platform several years earlier. The harasser has also targeted several of my closest online friends.

I believe the harasser may have stalked me across multiple platforms. For example, at one point the harasser referred to a GitHub link that I had briefly shared on my blog in response to a technical question—a link I had not shared elsewhere. This suggests that the person monitored my activity across multiple platforms. Nevertheless, Twitter remained the primary site where the harassment took place. One community member pointed out during member checking that this work does not address the cross-platform aspects of online harassment, which has been highlighted as an important area for research to understand harassment [9, 124]. Had I experienced harassment on other platforms with the same intensity, the system would have been designed differently.

Based on all available information, I assessed the likelihood of any physical safety risk to be low. I also do not intend to take legal action against my harasser, so preserving evidence is not a priority in the current tool.

4.1 Profile of the Target's Account

On Twitter, I have been using a pseudonymous account for over ten years. I do not use my real-world identity because in my home country people can be imprisoned for their Twitter commentaries. I have about 30 followers, most are online friends who do not know me in real life but have known me online for more than five years. Only three people I know in real life are aware of my Twitter identity. Even before the harassment campaign, knowing that a larger audience size is associated with increased harassment [126], I proactively removed followers whose identities were unclear to me in order to maintain a low follower count.

My Twitter account is used primarily as a personal journal. The engagement on the posts are very low naturally and are primarily from my friends. I do not usually discuss controversial topics with other users. My Twitter life had been largely quiet and pleasant before the harassment campaign.

4.2 Timeline of the Campaign

The harassment campaign, which has been ongoing for more than 9 months, is summarized in Fig. 2. It began suddenly one day in December, 2022 when an impersonating account, which appropriated my Twitter profile image, emerged out of nowhere and spammed my notifications with a string of offensive engagements. The account threatened to reveal my "scandals," attacked my appearance and intelligence, and threw out many baseless, random accusations.

My initial response was to block it and ask friends, both online and in real life, to help me report the harassing account and its tweets. However, on the second day, the harasser returned with another impersonating account. I once again asked my friends for assistance. The third day was quiet, but on the fourth day, the harasser came back with six newly registered accounts.

At that point, I realized that I needed to employ automation to deal with the harasser. Given the harasser's commitment, it seemed likely that the harassment would be long-term. I did not want to open my Twitter to see so much mental health-damaging content in my notification tab, nor did I want to go through these abusive accounts one by one, manually block them, and bother my friends again and again to report all the accounts and the tweets.

My instinct was correct. The harasser continued to return time and again. There were small waves of harassing accounts almost every month, and in June 2023, the intensity of the harassment reached its peak (see Fig. 2).

I have never directly engaged with the harasser's content. However, their unwanted interactions are visible to all who engage with my content, unless the harasser's accounts are restricted or suspended. Therefore, I typically delete "tainted" tweets (i.e., my posts which have harassing engagements) to prevent others from seeing the unwanted interactions I receive from my timeline. Once a malicious account has been acted on by the platform, its activity will be hidden by the platform. I keep this fact in mind when I design the automated reporting system.

The harassment campaign eventually came to an end in September 2023, ten months after the first message. With no explanation, the harasser stopped and has not come back since then.

4.3 The Harasser's Behavioral Patterns

4.3.1 Impersonation, Stalking, Bombarding, and Other Attacks. The harasser frequently steals my usernames[5] and uses offensive usernames embedded with my personal information. The harasser seems to enjoy role-playing and frequently impersonates both me (see Fig. 3) and accounts that have interacted with me.

The harasser makes use of all of Twitter's capabilities to make me read the harassing messages. The harasser posts myriad insults, threats, and leaks of my personal information via replies, retweets, and quoted retweets. Additionally, the harassing accounts attempt to follow me and my friends. My public twitter lists are also followed repetitively. My hypothesis is that the main purpose of this behavior is to make me and others read the insults embedded in the usernames.

My friends are also targets of the harassment campaign: the harasser's intention seems to be to also to make the harassment visible to my social circle. At times, this even involved spamming mentions at friends to draw their attention.

The harasser also used Twitter's protected account feature, which allows a user to interact with others without alerting them, for stalking and harassment. Engagements made by a protected account (follows, comments, etc.) are visible to its followers, but no notifications are sent to the recipients of these actions, which makes it convenient for a harasser to use an undetected protected account for surveillance and harassment purposes. The harasser attempted to use a protected main Twitter account (distinct from the harassment accounts), which I identified based on posts on another platform, to monitor my activity covertly. Additionally, the harasser attempted once to use a protected account with an offensive name to follow

[5] i.e., using my username on Twitter and other platforms (mostly home-country specific platforms) as well as creating lexicographically/typographically ambiguous clones or variations with offensive content added.

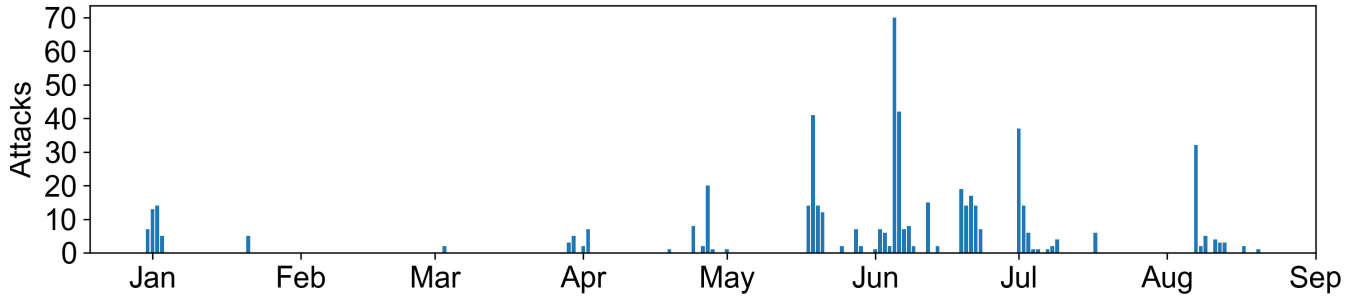


Figure 2: The harassment campaign escalated over several months from late 2022 with a peak in June 2023. "Attacks" refers to the harasser's comments, reposts, quotes, likes, follows.

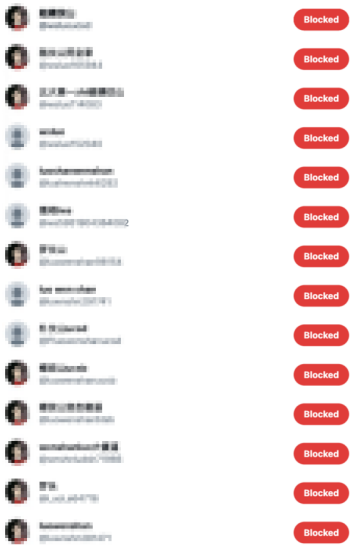


Figure 3: The harasser repeatedly used my profile picture and username in impersonating accounts. Pixelated screenshot from the Twitter block list interface taken in June 2023.

me, probably hoping that I would not discover it and remove it before people browsing my follower list had the chance to see it. However, in both cases, the harasser was caught and blocked immediately by my protection system. This was only possible because my system had the information from the account activity API.

The harasser is persistent. There are now 267 harassing accounts, 545 recorded unwanted direct engagements, and 610 additional posts posted after being blocked by me over the span of nine months. The temporal pattern of the harassment campaign is depicted in Fig. **??**.

4.3.2 Intermittent and Fleeting Attacks. The harasser uses new accounts for harassment shortly after creating them, usually within one hour of account creation. Half of the harassment starts within 2.5 min of account creation and 90% of the harassment starts within 7 min of account creation (see Fig. 5a). The consistent aggressiveness ironically makes it easier to identify the harasser's accounts automatically.

The harasser also posts in rapid bursts. Half of the harassing posts are posted with an inter-post interval of less than about 1 min and 90% of the harassing posts are posted with an inter-post interval of less than 1 hour 7 min (see Fig. 5b). This volume of activity from the harasser would inundate my entire notification tab if I had no protection during attacks.

The harassment activities occur intermittently. In total, the harasser harassed me on 56 days out of a nine-month period. While there are days without harassment, there are also days when the harasser created more than 20 new accounts. On the days when the harassment occurred, I received an average of approximately 10 unwanted engagements. The peak volume was 70 engagements on a single day. This intense but intermittent nature of attack makes it very difficult to predict when attacks will occur or to deal with them manually.

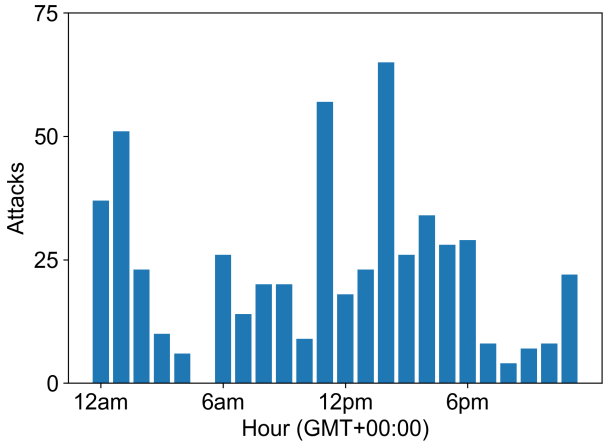
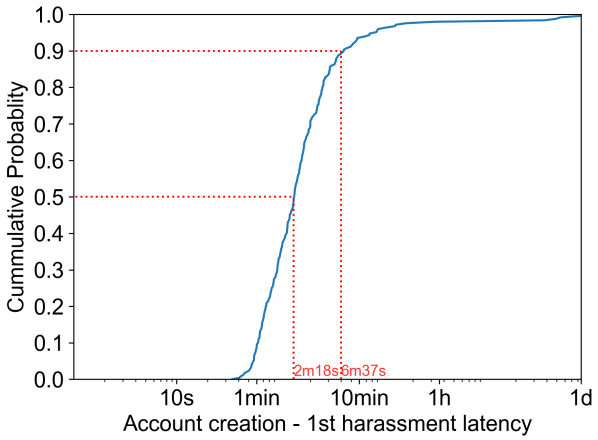
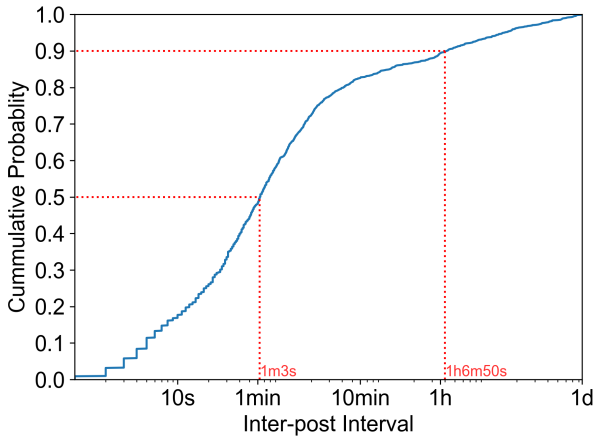


Figure 4: Aggregated count of attacks (the harasser's comments, reposts, quotes, likes, follows) per hour (timezone: GMT) throughout the harassment campaign. The harassment occurs at nearly all hours, even during my sleeping hours.



(a) CDF of harassment latency.



(b) CDF of inter-post interval.

Figure 5: The harasser initiates activity shortly after creating new accounts and posts rapidly (log scale is used for x axes). (a): The harasser starts harassment (comments, reposts, quotes, likes, follows) shortly after account creation. 50% of the harassment (lower red line) starts within 2 minutes and 18 seconds of account creation. (b): The harasser posts with low inter-post intervals. 50% of the harassing posts (lower red line) are posted with an inter-post interval of less than 1 minute and 3 seconds.

With my automated tool running, apart from the initial few days when I had no protection, all of the harasser's accounts were blocked within minutes of their first engagement with me (see Fig. 1 for an example). The harasser typically continued posting a few more times with each account after being blocked but eventually abandoned the blocked account and moved on.

I have a pinned tweet announcing that I am using automated tools for blocking. However, this has not deterred the harasser. The harasser simply continues to register new accounts for attacks. There is no evidence of automated mass account creation based on `created_at` timestamps from Twitter, and the harasser only registers a new batch of harassment accounts after accounts in the previous batch are blocked.

Twitter's rate limit did not affect the harassment behavior in any significant way. The harasser's newly registered accounts would not reach the rate limit before they are quickly abandoned.

This harassment campaign would qualify as cyberstalking, according to Dressing et al.'s criteria [32]. Even though cyberstalking is a less common form of online harassment [32, 94], its severity leaves considerable psychological trauma on victims [39].

5 The Personalized Anti-Harassment System

While targets of severe harassment often feel powerless and withdraw from digital space when the platform fails to protect them [13, 57], as a technical person accustomed to using various forms of task automation on a daily basis, I immediately recognized that I could employ automation to counter the harasser. After manually blocking and reporting the harasser's first ten or so accounts during the first wave of harassment, I decided to automate the process by building a personalized anti-harassment system.

5.1 Design Goals

Note that this list is of a retrospective nature. I initially developed the defense system in response to an ongoing harassment campaign, and it was not originally intended as a research project. Consequently, not all design goals were explicitly outlined at the beginning. My understanding of my needs evolved gradually as the harassment campaign unfolded. In its final form, the system aims to:

- Accurately identify interactions from the harasser.
- Mute and block the harasser promptly, given that the harasser posts very aggressively.
- Minimize my exposure to the various forms of attacks from the harasser. Missing all content from the harasser is not a concern for me, as the harasser does not pose any physical threat, and the insults have no merit.
- Remove the harassing content from the platform to the best of my ability, and minimize the spread of the offensive content within my social circles.
- Minimize time spent on reporting the harasser's accounts and posts.
- Minimize the harassment campaign's impact on my normal usage of the platform.
- Be flexible enough so that it could be used to filter other kinds of harassers.

5.2 System Overview

The protection system, which runs 24 hours a day, periodically filters the accounts that interact with the protected user by applying a general anti-spam rule and a stricter anti-harasser rule. It blocks the general spam accounts for the protected user and blocks the IDs identified as from the harasser for both the protected user and

the protected user's supporters. Finally, it saves those blocked IDs to a "harasser ID list."

The reporting subsystem regularly checks the "harasser ID list" and reports any accounts that are still unrestricted, based on actively probing the `UserByRestId` endpoint. To accomplish this, it utilizes the user's and the supporters' accounts, which are authenticated using cookies transmitted to the server via a dedicated secure website. The statuses of the reported tweets and profiles are tracked in a database to prevent duplicate reports.

The system currently runs on a dedicated server with almost no graphical user interface (with the exception of the cookies collection website). However, the system could be easily extended to include more traditional user interfaces.

5.3 Implementation Details

An system overview is shown in Fig. 6, and details of the major components follow in this section.

A **detection bot** examines the new engagements the protected user gets regularly by polling at a user-configurable frequency. Initially, this polled the `user_mentions` API when it was available; it now checks the notifications directly. It passes new engagements to the **judger** and then performs blocking and logging actions accordingly. Note that Twitter's `notifications` endpoint has a rate limit of 180 calls per 15 minutes. The client itself is checking the endpoint every 30 s, which means the user can still browse Twitter normally when the bot is checking notifications at a frequency as high as once per 10 seconds. Additionally, even though I have not used it against my own harasser, the bot also has the capability to conduct site-wide searches and take actions to address unwanted content that is not directly thrown at the victim. The `SearchTimeline` endpoint has a rate limit of 50 requests per 15 minutes, allowing for one search every 30 seconds. Optionally, the bot could be set up to send out a notification to my mobile phone via IFTTT to notify me that a new harassing account is detected.

A **webhook** (no longer freely available after the Twitter free API termination) receives unfiltered event streams from the user's account: if any interaction event is detected, the webhook starts the detection bot immediately (as opposed to waiting for a polling interval). A HTTPS domain running handlers for the user is required for using Twitter's webhook.

A **judger** filters the users interacting with my accounts with arbitrary, manual, heuristic rules that use account features such as account age, following count, followers count, statuses count, etc. In my case, since the harasser's behavioral pattern is highly unique (extremely short time between account creation and first interaction with me), no ML- or NLP-based filtering is needed to accurately identify the harasser. This is surprising, as ML and NLP approaches dominates the algorithmic content moderation work. But our experience highlights that behavioral patterns might be highly valuable for identifying dedicated harassment accounts in the case of severe, serial harassment. If the harasser were harder to identify, the judger could be extended with NLP-based filters or classifiers which examine the aggregated tweets, user names, and profile descriptions. There is also a custom allowlist which includes accounts I follow as well as my followers.

In the filtering rule, I can chain multiple sub-rules together so that it can be used to filter multiple types of accounts. Here is an example of custom filtering rule:

```
((followers_count <= 5 and following_count <= 5)
and (account_age <= 14)) or (account_age <= 1) or
(favorites_count/(account_age + 0.01) > 200)
```

This rule filters three types of accounts: new accounts with very few friends; even newer accounts with no restrictions on friend count; and accounts that are spamming favorites. I wrote a parser to handle arbitrarily complex rules so that there is no need to change the program logic. To tune the filtering rules, I first calculated the statistics of key attributes for the harassing accounts and compared them with the attributes of the accounts I follow and those that follow me. I then set initial thresholds based on these distributions and iteratively adjusted them to minimize both false positives and false negatives.

A **reporting bot** runs every few minutes. This component uses my own and my supporters' accounts to report every unreported tweet from accounts on the "harasser list" which are still in "normal" state, until they are restricted. It is not sufficient for the reporting task to be a one-shot effort upon detection because there is a possibility that the harasser may create additional posts after being reported for the first time.

Twitter offers a large number of, in my opinion, ambiguous flags for reporting content. Some researchers have argued that this is an intentional move to avoid responsibility and suppress disagreement [25]. To address both the uncertainty and the difficulty of categorizing violations under Twitter's content policy, I designed the system so that users can configure the bot to always report an account or a tweet using multiple flags, thereby increasing the likelihood of success. It should be the responsibility of the platform, rather than the user, to verify the accuracy of the flags. To make reporting with multiple accounts and multiple flags faster, all steps required for completing a report are implemented as asynchronous HTTP requests. It takes about two seconds to complete the reporting for a single flag with ten friendly accounts.

A **logging system** records all interactions made to my Twitter account and maintains a comprehensive record of all reports. It logs the accounts that have been reported, the posts that have been reported, and the current status of the reported accounts, which can be categorized as removed (suspended or deleted), normal, and restricted. Additionally, it keeps track of the statuses of my posts that have received engagements from the harasser, which indicate whether they have been preserved or already deleted by myself. This way I am able to balance damage mitigation and the goal of keeping my account running as normally as possible.

To prevent exposure to harassment injected in usernames, in the logs that I check regularly, the harasser's accounts are represented by their numerical IDs only.

A **plotter** generates plots and statistics from the log files. The simple visualization allows me to view the harasser's behavior pattern

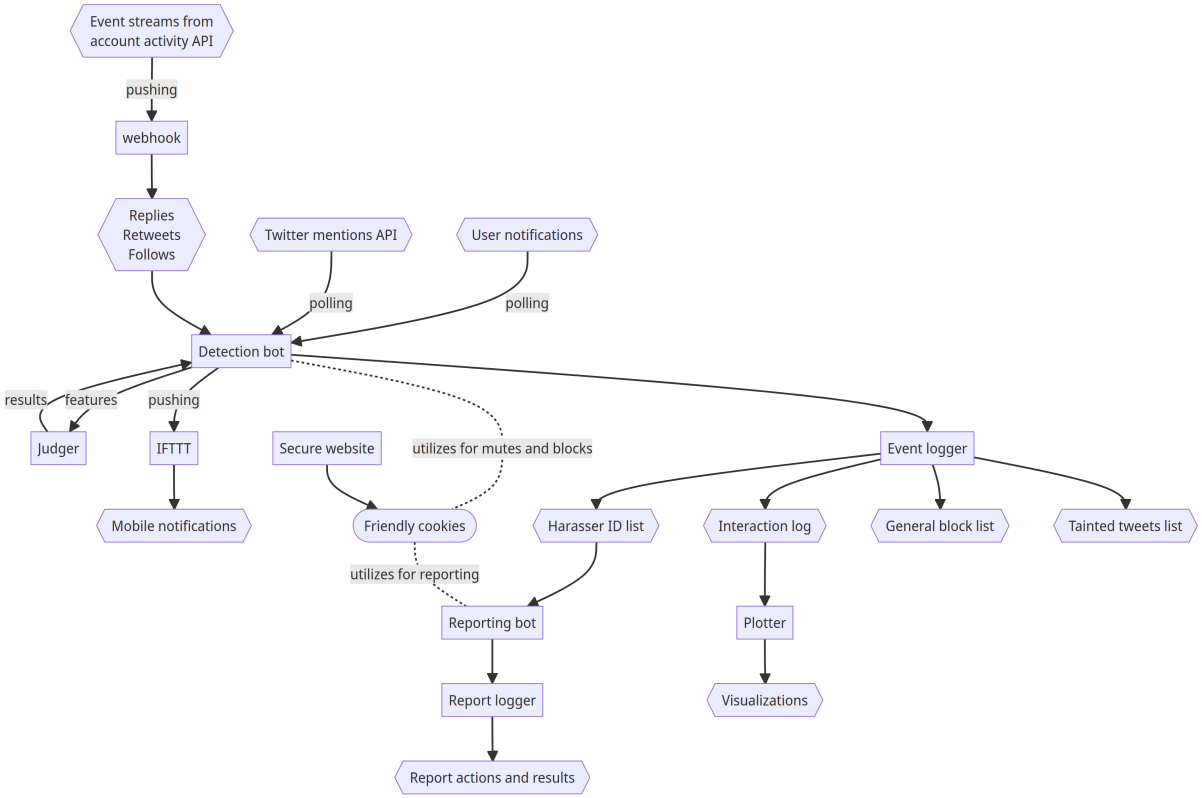


Figure 6: Overview of the system and its major components.

easier.

Finally, a simple **website** makes it easier for my friends to directly deliver their cookies to my server, for coordinated blocking and reporting. The website also provides instructions for installing trusted open-source browser extensions for exporting cookies.

5.4 The Evolution of the System

As the developer of my own defense system, I am responsible for adapting to any changes. Although I initially assumed the development process would be straightforward, my system soon began to break down as the official APIs I relied on became inaccessible due to Twitter's policy changes. As I describe later, working with the official API is comparatively manageable, as it is at least documented. Once the official API became unavailable and I had to rely on reverse-engineering, however, the workload increased substantially. I was forced to rebuild the system repeatedly in response to sudden changes, which created significant technical and emotional challenges. Those challanges are also integral to my autoethnography.

5.4.1 API Days. I began building the system a few days after the harassment started in December 2022. At that time, the Twitter APIs were still freely accessible. The free 'essential tier' had a monthly usage cap of 500,000, while the free 'elevated tier' had a monthly usage cap of 2,000,000 and also provided access to the account activity API. Since the account activity API was necessary to directly obtain the list of quoted retweets and retweets made to my account, I applied for elevated access immediately and received approval on January 2, 2023.

Twitter API v1.1 offered strong support for my anti-harassment system. An account activity API provided unfiltered access to all account activity in real time. This included follows, replies, retweets, quoted retweets, and more. All relevant metadata was included in the notification, which eliminated the need for additional API calls. To receive notifications from this API, I had to register my own webhook. Based on my experience, this was the most valuable feature for capturing harassers in real time. Additionally, I could even receive notifications about engagements from protected accounts (the harasser could abuse protected accounts for stalking and harassment: I did catch the harasser's protected accounts that tried to follow me via the account activity API), a feature missing from Twitter's other APIs and endpoints. With the help of the account activity API, the latency of blocking could be as as low as five seconds.

In contrast, Twitter's official API v2 posed challenges for building anti-harassment tools. It lacked a unified API to retrieve all recent interactions with a user in one request. The `mentions` endpoint only provided mentions and replies, which excludes quoted/unquoted retweets (which my harasser frequently engages in). To retrieve all retweets of my tweets, I had to use the `retweeted_by` and the

`quote_tweets` APIs to examine each tweet individually, because there was no API that would provide me with a list of tweets that recently received retweets. Unfortunately, this is impractical when I have more than 10,000 tweets and the harasser might retweet any of them. Furthermore, the results delivered by API v2 are subjected to platform-wise anti-spam filtering and the user's own quality filters. With access to only this filtered event stream, the protection tool might miss harasser engagements and thus fail to mitigate their effects.

5.4.2 The End of the Free Twitter API. On February 2, 2023, Twitter made a sudden announcement that the free access to Twitter APIs would soon come to an end (although I still had free access to the account activity API until early June 2023). The new APIs were significantly limited compared to the previous offerings and were priced in a way that made them unaffordable for individuals. The account activity API, which was very useful for fighting harassment, was locked behind a $42,000/month price tag. This effectively meant the end of my existing anti-harassment system, as well as the demise of other third-party anti-harassment tools [136].

Determined to continue combating the harasser for as long as I remained on the platform, I started modifying my system to eliminate the need for the official Twitter APIs. I accomplished this by reverse engineering the Twitter web client through a careful study of its HTTP requests. Since some parameters are generated on the client side, I studied the obfuscated JavaScript code and used trial and error to determine how to generate tokens and IDs, such as `report_flow_id` and `x-client-transaction-id`, required in the HTTP form to submit valid requests. This required learning new skills for reverse engineering online platforms.

Since the reverse-engineered endpoints lack official documentation, I have to rely on experimentation to determine their rate limits. Surprisingly, most of the endpoints used by the Twitter client itself have fewer restrictions than their API counterparts. For example, the paid Pro version[6] of the API rate limits the number of blocks per user to 50 per 15 minutes. However, the endpoint used by the client has no problem blocking 100 accounts without stops in my testing. The theoretical minimal latency for blocking allowed by this approach is 6 seconds given the rate limit of the `notifications` endpoint. In practice, I set the polling frequency to once per 30 seconds, which is sufficient for my protection.

I have experienced several emergency situations caused by Twitter's sudden changes to endpoints after switching to reverse engineering. In early June 2023, Twitter unexpectedly changed the login flow. At the end of June 2023, Twitter made authentication mandatory for `UserTweetsAndReplies` and `TweetDetail` endpoints. There were also a few instances where an endpoint suddenly started to require a new field related to the newly added Twitter monetization scheme in the HTTP request payload. Each time, these changes broke my system as a result. Despite the fact that, in each case, I was able to fix the problem with little effort, I did feel vulnerable during the update process and worried that the harasser would exploit the opportunity.

5.4.3 Coordinated Anti-harassment. In May 2023, the harasser began another wave of aggressive harassment. This time, the harasser also spammed mentions of my friends. That is when I decided to gather my friends' cookies so that we could block and report the harasser together.

[6]Now priced at $5,000/month [131]

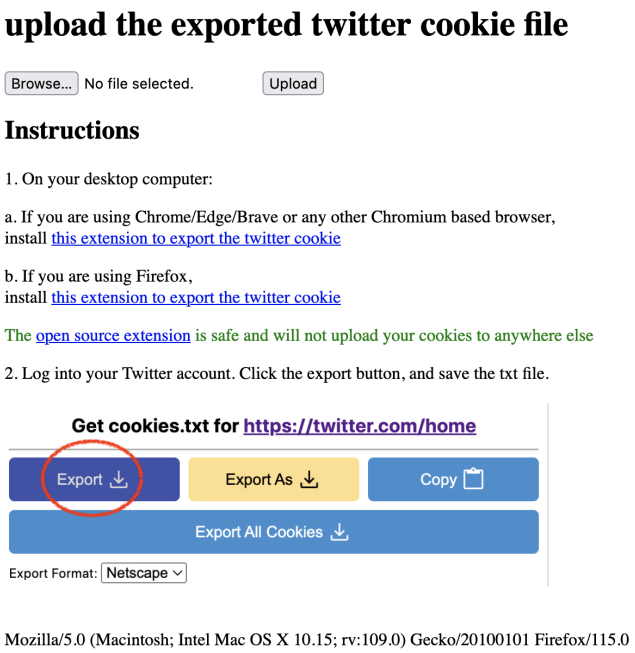


Figure 7: To ease participation, instructions for collaborators are given on the cookies collection website, with an upload feature for delivering session cookies as revocable credentials directly to my server running the anti-harassment system.

I created a Twitter chat group and added my friends to it. I explained my idea to them and made them aware of the risks and downsides involved. All of my friends agreed to assist me, and they trusted me enough to provide access to their session cookies. The use of this long-lived, ephemeral identifier—somewhat akin to an application-specific password in other authentication contexts—derisked credential sharing for my friends. Initially, I asked friends to deposit session cookies using a cloud-based storage platform. However, after one friend mentioned that she had not used this platform, I decided further ease the process by making a dedicated website, shown in Fig. 7, for my friends to simply upload their session cookies.

I was compelled to use cookies as an authentication method because the free Twitter APIs are no longer accessible. Twitter's session cookies remain valid as long as the user does not log out from the session associated with the cookies. Unfortunately, this method granted me more privileges over my friends' accounts than necessary. It also imposed additional burdens on my friends. Instead of a simple click, they now have to open their desktop browser, install an extension, export the cookies, and send the highly sensitive cookies to me. In the past, when free Twitter APIs were still available, I could block users on behalf of my friends by obtaining their authorizations through OAuth, which only requires one click.

Participating in automated reporting also leads to my friends' Twitter notification tabs being flooded with reporting-related notifications whenever the harasser shows up. I am grateful that my friends offer me their full support and are understanding of these inconveniences. However, it would be helpful to have an option to mute the reporting-related notifications from the badge counter without hiding them from the notification tab completely. Additionally, having a notification tab where all reports related to a single account are collapsed together would minimize the impact on the user experience of my supporters.

Ultimately, the attacker clearly wants to harm me socially. Therefore, if none of the people who know and care about me see their offensive content, the attacker's efforts are pointless. Even though I am transparent about my use of automation, I have not disclosed that I am blocking for my friends. The harasser, not knowing this, has actually thrown "Why are you so shameless when I expose you?" at me. My ability to extend protection beyond my direct view of the platform to encompass the experience of my friends has been a critical facet to minimizing the actual harm caused by these attacks.

5.5 The Demonstrated Performance of the System

Before the system was deployed, my Twitter notifications were frequently flooded with messages from the harasser, and ever since the system was deployed, I have seen no content from the harasser through Twitter's non-report notifications. In every instance, the harasser's accounts were blocked before I had a chance to review the notifications. On average, the harasser's accounts have only 2.0 engagements with me before they are blocked. All of the harasser's accounts were identified and blocked, with only one case of misidentification.[7]

As a result, I am able to continue to use Twitter.

As of July 15 2023, 81% of the 231 accounts belonging to the harasser have been suspended or deleted as a result of automated reporting. As of October 11 2023, 96% of the 267 accounts belonging to the harasser have been suspended or deleted as a result of automated reporting. Because I did not conduct controlled comparisons between manual reporting and reporting through my system, I cannot determine whether these accounts would have been suspended through normal manual reporting without the system's involvement. However, this suspension rate is notably high comparing to published reporting statistics. In a study conducted by Woman Action Media (WAM!) in partnership with Twitter [89], the suspension rate for accounts reported in escalated reporting tickets was only 55%.

A total of 81,048 reports were submitted, which is the result of all friendly accounts reporting every tweet of every account of the harasser, and each tweet is reported using multiple categories. It would take 67.54 person hours to submit the same number of reports manually, assuming each report takes 3 seconds. Given the difficulty of taking down abusive accounts, this approach is unfortunately necessary. However, much of the reporting could be avoided if Twitter permitted users to remove harassing comments directly, a feature available on many of the other platforms discussed in Appendix A. It is quite possible that our high volume of reports added to the burden of Twitter's already overworked moderators, many of whom are likely low-wage workers based in the Global South [106]. At the same time, it is also possible that there was little human review involved at all, and that our reports merely fed into an opaque algorithmic moderation process. Either way, I attribute the high suspension rate I observed to consistent and concerted reporting efforts. The median time for Twitter to take actions on the reported accounts is 14 days, and the statuses of the harasser's accounts at July 15, 2023 and October 11, 2023 are shown in Fig. 8. Without exception, based on the notifications I receive from Twitter, the harasser's accounts are always found to be violating the rules regarding hateful conduct, abusive behavior, and exposing private information simultaneously.

The system designed in this work mobilizes friends, and by leveraging automation, it minimizes the effort required from them. Beyond the initial upload of their Twitter cookies, no further action was needed.

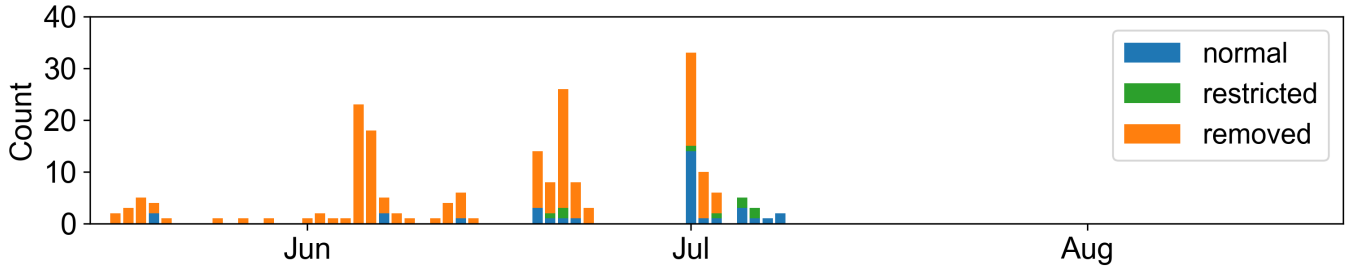
5.6 Reflections on System Design

5.6.1 Considerations of Making the Current System Available to Other Harassment Victims. The current system is effective in automatically dealing with my harasser, who is a single obsessive stalker using a series of dedicated harassment accounts to launch sporadic attacks.

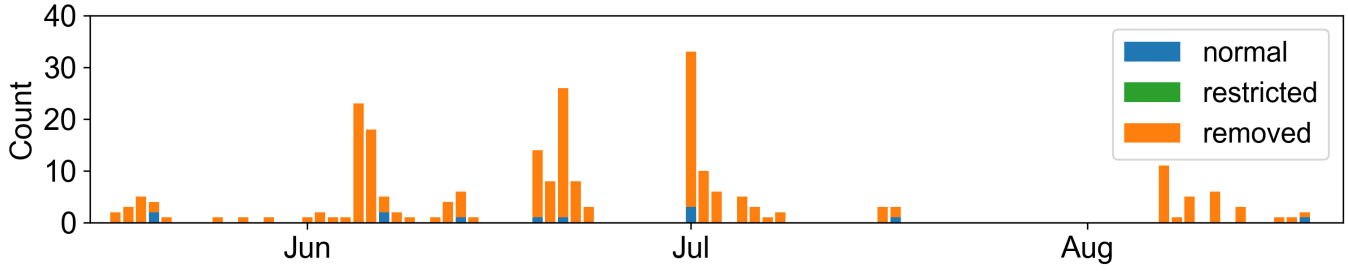
I believe an approach heavily leaning towards automation would be particularly effective against stalkers/harassers who use dedicated accounts for the purpose of targeted harassment. This is because their accounts often display unique patterns of speech and behavior. It has been shown that textual forensics techniques could be used to uniquely identify the originator of cyber-stalking [40]. In my case, 99% of my harasser's accounts start harassment within one day of account creation, and 66% of their screen names contain identifiers of me or self-references of the harasser. Paid harassers tend to adhere to specific scripts and recycle the same sentences repeatedly. Using my tool, in December 2023, I discovered a new Twitter harassment campaign targeting a Chinese user residing in Italy who played a prominent role in the 2022 protest in China. It's likely that the campaign is state-sponsored as the accounts' behavioral patterns are consistent with Chinese nation-state-sponsored trolls reported previously [132]. Out of 100 harassing posts and replies I gathered, only 10 unique sentences are found. Whether unpaid or paid, another hallmark of dedicated harassing accounts is that these accounts do not have normal content expected from a typical account: such accounts mostly post harassment, rarely talking about normal life or interests.

During member checking, one person noted that the current system design does not address networked or coordinated harassment campaigns. Technically, the system is flexible enough to accommodate various verbal and behavioral patterns. However, addressing casual toxic comments from mobs that are not fixated on any specific victim and are not full-time trolls would be a more complex. This is because, as noted by Cheng et al., "anyone can become a

[7] The misidentified account was a newly registered, non-malicious account with no prior posting history that liked one of my posts. It was blocked immediately. Since the account did not post any harassing posts later, I concluded that it had not been from the harasser.



(a) As of July 15, 2023



(b) As of October 11, 2023

Figure 8: Account status by account creation date. Only the accounts created after May 15, 2023 are included, as all accounts created earlier have been suspended at both dates. Typically, account removals take two weeks, although a very small number of accounts (4%) were never removed.

troll" [22], and their accounts might lack the distinct hallmarks that could be reliably recognized by a machine.

As mentioned in Section 5.1, since I assessed the harasser to pose no physical threat, I chose to ignore the harassing content altogether and did not preserve it. This may not hold for other targets of severe harassment, who may need to examine abusive messages in order to assess potential real-life threats. NLP-based techniques could help reduce the emotional burden of such threat assessment, for example by replacing entities in abusive messages while preserving their semantics [67]. However, deploying NLP models requires careful attention to privacy, as harassment content may contain sensitive personal information.

5.6.2 Would the Approach Work with Users without Strong Ties on Social Media? I am fortunate enough to have friends who know me and are very supportive to make their accounts readily available. Squadbox [86] showed similarly that, with the right infrastructure, enrolling friends for peer-to-peer content moderation to address online harassment is possible. A similar infrastructure for account sharing could benefit users with a limited number of close friends on social media, providing them with additional support.

However, prior work has also shown that strangers can be used to accomplish similar ends. HeartMob [13] illustrates the feasibility of recruiting volunteer strangers to aid in reporting. Blockbots have been used in similar ways to support collective action among strangers [43]. Anonymous users who do not want to share their real-world personal details, may hesitate to seek help from online friends when dealing with a doxxing harasser. In such cases, obtaining assistance from stranger volunteers would be more suitable. In summary, friends and strangers can be recruited to help fight harassment together. The most suitable approach varies based on the user's support network strength and privacy considerations.

5.6.3 How to Make the System Available to Non-Technical Users? At present, the system lacks a graphical user interface as it is exclusively used by myself, and I am comfortable logging into the remote server where the system is running to execute commands directly. However, this does limit its accessibility to a non-technical audience. The system could be set up by anyone with access to servers and GUIs could be potentially developed to make it easier for non-programmers. The process of tuning the filtering rules could be further automated, and an interactive interface that provides instant feedback on the effects of filtering would help non-technical users configure their filters more easily.

To extend consistent protections to a larger user base, computing resources beyond a single server are needed and the amount of maintenance work is more than what a single person can do. If not integrated into the platforms themselves, to keep the system free, decentralized deployment by volunteers is the only viable approach. Block Together [60], which ran as a centralized service, eventually closed because the site owner could no longer get the resources to maintain its operation for 300K users. Ideally, the system could be integrated into the platforms themselves, as only the platforms possess the resources to scale easily. However, the experience of Block Together, where Twitter implemented a much less usable version of the blocklist feature without automatic updates despite advice from the developer, suggests that the platform may not always prioritize quality when integrating features initially developed by users [60].

5.6.4 Would Decentralized Social Media Make My Life Easier? Recently, decentralized social media, represented by the fediverse, has

emerged as an alternative to centralized, commercial social media platforms [81, 104]. There is hope that decentralization would help address the scalability problem of content moderation inherent in centralized platforms [18].

However, the harassment campaign I experienced on Twitter is still possible on federated social media and is not necessarily easier to address. First, since user registration is managed by each instance, the harasser can use the same email address to register harassing accounts on multiple instances, eliminating the need to create many email addresses before starting the harassment campaign. Second, since the nature of the fediverse requires collaboration between instances, it is possible for bad faith actors to set up non-compliant instances that implement the protocol differently to bypass some effects of the content moderation actions. For example, an instance that does not implement blocking in compliance with the protocol could deliver your post to the blocked user against your will. Finally, there is no guarantee that the instance administrators will handle reports quickly. Some administrators may not care and others – even if they are willing to cooperate – might already be overwhelmed by moderation tasks and respond slowly [4].

Federation also makes reasoning about and implementing features related to content visibility more complicated (see [68] for the flowchart that determines a reply's visibility). Currently, as shown in Table 1, Mastodon's implementation of blocking does not address third party visibility. On the plus side, the open nature of the fediverse ecosystem guarantees that as a developer, the API will always be available for free, not behind a hefty price tag.

5.7 Further Reflections

5.7.1 My Alternatives to Building the System. **Why I didn't simply lock my account.** I did not lock my account because I found it to be too restrictive and ineffective in the face of a committed harasser. A temporary lock will not help: as soon as I drop my guard, the harasser will return. A persistent lock means the victory of the harasser: their verbal violence has successfully pushed me out of the public space on the platform.

Why I didn't simply abandon my account. Changing to another account is another option. However, since the harasser has been stalking me for an extended period of time (some of my followers told me that it was the harasser who identified and revealed my Twitter account to them in the first place, while I have never personally told my online whereabouts to her), my frequent contacts on the platform are already known to her. As a result, as long as I keep connection with my friends in the new account, the harasser will uncover my identity eventually.

Why I didn't just persuade the harasser to stop. Over the harassment campaign I asked mutual friends multiple times to encourage the harasser to seek therapy instead of wasting time harassing me. Unfortunately it did not work.

Why I didn't just leave for another platform. I have tried multiple Twitter alternatives, but the majority of the accounts I am following still post on Twitter. Because of the strong network effects [35], I find it hard to leave Twitter.

5.7.2 Was I Fixing Twitter's Problems for Them? The problem-solving approach to content moderation has been questioned by scholars. Gillespie recently argues that we need not fix industry-made problems on their terms [48]. This raises an uneasy question: by building the self-defense system, was I simply doing Twitter's work for them?

On one hand, my system filled gaps left by the platform's inadequate safety tools to deal with persistent harassers. On the other hand, my aim was not to improve Twitter's business model or help capture more revenue but to secure safe environment for myself and other victims of targeted harassment. The tension between my interest in safety and the platform's interest in profitability created the very conditions that pushed me to rely on reverse-engineering. The need to build my own tools was evidence of a structural misalignment that better platform design could have eased, though never fully resolved.

6 Limitations, Flaws, and Abuse of Twitter Protections, and Recommendations to Other Platforms Based on Those Lessons

The existing anti-harassment tools provided by Twitter are quite limited and its "protective" features can actually be exploited for harassment.

Because of our frustrations with Twitter, we have a series of recommendations for the designers and developers at Twitter and similar platforms. Although we primarily illustrate issues using Twitter, our discussion considers platforms beyond Twitter, and most of our claims are supported by related work. As argued in [95], content moderation involves both easy and hard problems. Although policymaking remains hard, we believe that addressing a sustained, targeted harassment campaign is possible.

6.1 Existing Policies Are Not Enforced

6.1.1 Issue: Mass Account Creation and Ban Evasion. Account creation is the first step in the harassment workflow. How can a single person register more than 200 Twitter accounts while repeatedly being suspended? Theoretically, Twitter has an "Inauthentic Activity" policy which disallows mass account generation from a single person.[8] Similarly, Twitter claims suspended users are prevented from registering again, in the account suspension notice, but this did not seem to prevent my harasser from successfully registering hundreds of accounts. It is possible that the harasser circumvented restrictions by employing techniques such as registering on different physical machines or using VPNs to obtain new IP addresses. However, in my own experiments, I have found it is possible to register dozens of accounts from a single machine simply by using different email addresses—no phone number is required for registration, no block on the IP address ever occurred. This may be due to bugs in their code,[9] a focus on spam and/or coordinated activity that fails to catch "small fish", or simply a one-off exception. Without access to Twitter's internal data, it is impossible to know how many (if any) accounts are caught. Whatever the cause, my experience highlights the consequences of a breakdown between policy and enforcement, which researchers have extensively discussed. Succinctly, policies and guidelines are not uniformly enforced, and

[8] https://help.x.com/en/rules-and-policies/authenticity

[9] As when Facebook blamed a bug in their API for exposing user photos to third-party app developers [8].

double standards abound [30, 54]. Further, the enforcement process is typically opaque, confusing to users [123], and resistant to external scrutiny [75]. In my case, I have no idea how Twitter handles mass account creation or whether there is a prevention mechanism in place at all.

6.1.2 Recommendation: Improving Fake Account and Ban-Evasion Detection. The kind of severe harassment described in this paper is only possible because a single committed person can register hundreds of accounts. On platforms where anonymity is prioritized, it may be legitimate for a single entity to register multiple anonymous accounts. Users could leverage this anonymity for legitimate uses such as sharing sensitive experiences and seeking support [3]. For such platforms, identity verification-based prevention might not be an option due to privacy concerns. However, even in such cases, actions can still be taken after a large number of malicious accounts are created. Currently there is a rich literature on fake account detection in social networks, with most techniques categorized as either content-based or social graph-based [77, 103, 107]. Social graph-based methods rely on the assumption that fake accounts tend to be connected to other fake accounts, and that real accounts tend to be connected to other real accounts [17]. In my case, the harasser tends to create networks for fake accounts, likely in an attempt to game Twitter's account-quality evaluation system in which account metrics might play a role, to avoid automatic restrictions on low quality accounts. From the platform's perspective, verifying the connections between low-quality accounts with a shared history of harassing a specific target is relatively easy. Platforms could even treat known victims differently, restricting any new accounts that contact them once the victim has reported an initial harassment campaign.

Ban-evasion detection in the context of online harassment is also under-addressed. Evasion accounts often exhibit behavior similar to their banned parent accounts, as observed in previous research [97, 128]. However, some behavioral patterns—such as the brief interval between account registration and the onset of harassment I experienced—have not yet been studied in the literature, and could be leveraged for the rapid identification and removal of evasion accounts.

6.2 Visibility Controls Are Limited and Difficult to Understand

6.2.1 Issue: Algorithmic Visibility Rules Are Opaque and Inconsistent. While there are algorithms governing the visibility of comments on the site, the impacts of different actions are not shared with users. Sometimes a harassing comment is labeled offensive and hidden behind a banner automatically. Sometimes it is completely hidden, without any banner. However, the rules are not transparent and in many cases the harasser's comment is left visible to everyone. The inscrutability of Twitter's algorithmic moderation often makes me wonder about what has happened and what will happen next. Is that hidden comment (with only the comment count visible) from the harasser or from a regular spam bot? Is it hidden temporarily, or will the hidden content reappear after the restriction put on the author expires? Therefore, in addition to the issues of transparency, fairness, and depoliticisation raised in [50], algorithmic content moderation could also deprive the harassment victims the necessary information to fight back effectively.

6.2.2 Recommendation: Transparent and Consistent Algorithmic Visibility Rules. Platforms should be transparent about how different abnormal account states affect the visibility of posts and whether those effects are reversible. When engagements are hidden, platforms should indicate this clearly and show the timeline or duration of the hiding effect in the engagement statistics.

6.2.3 Issue: Limited Options for Making the Harassment Disappear. For the end user, there is no way other than reporting to make the harassment contents disappear from the public view. Unwanted comments and quoted retweets cannot be hidden. Twitter's reply-hiding function is designed so that if you hide a comment, a large, noticeable banner appears in its place. In my experience, this arouses the curiosity of bystanders. Additionally, quoted retweets cannot be hidden at all.

Most disappointingly, the blocking function is not designed to address the social aspect of harassment. Even after I block an abusive account, its unwanted interactions are still visible to my followers and other people who are unaware of the harassing situation at all, making it ineffective to address doxxing attacks.

As discussed in Section 2.2, blocking can be made collaborative through the use of shared blocklists [43, 71]. While blocking together could reduce the visibility of harassment among the subscribers of a shared blocklist, its effectiveness is still limited by what a block can do. On Twitter, as long as the harassment remains on the platform, it is still visible to anyone who does not participate in the shared blocklist.

Finally, there is no way to delete unwanted responses to my posts completely. Even if I delete the original post to remove any traces of the interaction from my own timeline, the replies and quoted retweets still remain on the platform and often continue to appear in search results. Twitter's design choice of not allowing users to delete comments directly is deeply frustrating. On my self-hosted website, I can freely remove unwanted replies without anyone questioning my right to do so. On Twitter, by contrast, the scope of my control is defined by the platform, whose interests may not always align with mine. As Srnicek notes, "In their position as an intermediary, platforms gain not only access to more data but also control and governance over the rules of the game" [118]. As a result, I have to rely on reporting as the primary mechanism of harassment removal.

6.2.4 Recommendation: Smarter Visibility Rules. My experience has also shown many of the shortfalls in how platforms make decisions around visibility. Social media users frequently reflect on their audience due to self-presentation considerations and privacy concerns, and the platforms have introduced features that allow users to tailor how they present themselves to different audiences [28, 64, 138]. However, the user is not the only one who can disclose their information; other users may also share information about the poster, whether they like it or not. Even algorithmic systems can expose private information to unintended audiences [137].

When a harasser targets someone openly, the target is rarely the only intended audience. Sometimes the harasser deliberately weaponizes content leakage to harm the target [5]. Some researchers

place a low priority on content leakage, because it is less prevalent and out of the control of the target [142]. However, as a victim of this type of attack, I encourage platforms to consider the social dynamics of online harassment when designing visibility rules for content moderation tools.

Visibility is an important aspect of moderation activity. Many forms of online abuse, doxxing and impersonation in particular, are intended to be seen by an audience [125]. And researchers and practitioners have explored the role of reduced visibility in curbing negative behaviors [47].

All suggest that platforms need more care in designing moderation visibility outcomes. For example, if user A blocks user B, there should be an option to hide B's quotes, replies, and retweets on A's page. This measure would help reduce the social harm of harassment. In my case, I obtained my friends' cookies to perform blocking for them, but this should not be necessary.

6.3 Mutes and Blocks Are Inadequate

6.3.1 Issue: Reactive Moderation Does Not Prevent Harassment . Twitter provides options to block and mute. Both options require me to read the harassing posts before performing a block and/or mute through my notifications—exposing me to precisely the content I wish to avoid.

The notification settings offer a few options for automated mutes (under the "Quality Filter"), but I felt they did not provide sufficient control. For example, I still want to be able to interact with friendly strangers. However, the only available options are muting notifications from people who do not follow me and muting notifications from people that I do not follow. At the time, I considered using BlockParty [15], which would give me more control, but it still would not allow me to precisely filter interactions from my particular harasser.

6.3.2 Recommendation: Enhance Proactive Moderation. Currently on most platforms, users can reactively block and report accounts, after experiencing unwanted interactions. To prevent severe and repeated harassment, it is crucial to implement moderation mechanisms that deal with harassment proactively. This can both help victims and reduce the total number of reports platforms receive.

Such mechanisms could be implemented at different stages of online interactions. For example, block lists could be used to disallow interaction initiation from certain users, although they cannot handle accounts that have not been encountered [43, 71]. To deal with unknown accounts, configurable rule-based filtering might be necessary. Currently you can set who can engage with individual posts on platforms such as Twitter, Threads and Bluesky; but those settings are limited to "allowing interactions from people you follow," "allowing interactions from people you mention," and "everyone." This does not give targeted users the kind of granularity they need for normal platform usage. A better system might allow targeted users to more precisely control the proactive filtering strategies. For instance, I may wish to disallow interactions from accounts less than one year old, unless I have initiated interactions with them in the past. To help users avoid overly limiting filters, platforms could provide estimates on how many interactions would likely be affected based on their past interactions.

At the stage of post creation, some researchers have tried to use real-time feedback to encourage the users to adhere to Community Guidelines, and it has shown promise in reducing toxicity [115, 125]. After a post is created, there are still opportunities to prevent it from reaching the intended target. Some platforms offer custom muted words, allowing users to ignore unwanted interactions. However, these filtered interactions are often publicly visible by default. To make this feature more helpful for targeted accounts, muted words could be categorized into two groups: those that prevent interactions from appearing publicly by default and those that simply hide them from the user. Finally, manual pre-screening, with which all content must be approved by moderators before being displayed, has been used by some websites [78]. However, requiring users to approve every interaction, while partially mitigates the harms of harassment, still forces the user to read harmful content from the abuser. At this stage, a combination of algorithmic filtering and an option of manual inspection could be more effective than fully manual pre-screening. Algorithmic proactive moderation has been suggested to reduce harm in a scalable way [109], and users may like to shape the algorithmic filtering in ways that suit their particular needs. While researchers have identified limitations of large numbers of control settings (as in the case of setting custom muted words) on social media platforms [63], these concerns are largely directed at the general public. We discuss the unique needs and motivations of targets of harassment in Section 6.6.

A more proactive approach to moderation is not without tradeoffs. One community member we consulted during the member checking process noted that automated filtering must be used with caution, as it can unintentionally suppress marginalized voices. One can also imagine public figures using automated filtering to silence criticism. For example, if account-age–based filtering is available, a politician could purchase long-established accounts, coordinate them to amplify their own voice, and block the majority of users with newer accounts. As discussed in Section 6.6.2, designers must balance public interests and user safety when deploying moderation features. Some tradeoffs may be justified for individuals facing severe harassment, but the same approach may not be advisable when applied to a broader user base, particularly public figures.

6.3.3 No Platform Has Done It Right for Me. Even though Twitter is particularly bad, surprisingly, no platform that I have regularly used has implemented account-level moderation that addresses the case of an obsessive harasser. Using a series of accounts and experiments on types of interactions, I have tabulated the attributes of platform moderation functions (Table 1 in Appendix A). These tests were needed because help pages and user guides usually do not specify sufficient detail. I group the results into four categories: interactivity, visibility, scope, and alert, to characterize how I, my network, and the attacker are impacted by the moderation action.

Ideally, the implementation should limit action-issuer initiated interactions (to prevent abuse by the harasser for harassment that is undiscoverable or unaddressable). It should take third-party visibility problems into account, apply to all past and future interactions, apply to other accounts created by the same user, and be reversible (to prevent communication loss in case of mistake).

Some features of my ideal blocking implementation may add complexity to the platform's system design. For instance, extending

a block to all accounts created by the same user is only partially feasible unless the platform implements a system to identify linked accounts. What I would like to address here is that the platforms' implementation of seemingly universal moderation features, such as blocking, is surprisingly varied, which can be quite confusing for regular users.

6.4 Privacy Features Support Further Attacks

6.4.1 Issue: My Harasser Abused Privacy Features. As mentioned earlier, my harasser used privacy features provided by Twitter for harassment. The interactions from a protected account *do* appear in the raw event stream available from the account activity API, but there is no way for a regular user to know from the notification tab. This also enables several new forms of attack. A simple attack is to follow the victim using a protected account with offensive user names, and the victim would not discover this unless they check their follower list. More severe attacks are also possible. The harasser could use a protected account to leave a large amount of harassment, then set the account to public, making all the harassment visible. But because the account was protected when the content was initially posted, the victim would receive no notifications for the harassment that occurred.

Another related attack is to use Twitter's private list feature to add harassment targets, then make the list public (again, no notification). This list could be shared to launch a concerted harassment campaign that is entirely unexpected by the victims on the list. I have not experienced either attack, but after my first experience of a protected account attack, I began to investigate Twitter's notification policies. Through this exploration, I discovered this private-to-public design flaw that could be exploited by a smart harasser who has studied Twitter notifications.

Other sites which implement similar private account features tend to have similar vulnerabilities. For example, I have found that Threads have the exactly same notification issue that makes stealth attacks possible.

Surprisingly, blocking could also be abused by the harasser. On Twitter, blocking is not entirely symmetric: after you block someone, you can opt to view their posts and engage with them. They will not receive any notifications about your engagements and your engagements are invisible to them but are visible to third parties. The harasser thus can block the victim first before attacking.

According to Strohmayer et al., "designers have a responsibility to anticipate and mitigate the possibility that malicious actors might weaponise a system for harmful activity" [121]. Privacy features are no exception, even if they were originally designed for user protection.

6.4.2 Recommendation: Consider the Potential for Abuse When Designing Privacy Features. Designs originally intended to protect user privacy can be utilized by harassers to secretly stalk other individuals without alerting them, which also enables the launch of surprise harassment campaigns.

Previous literature on stalking has mentioned the use of spyware and stalkerware—covert, malicious software that exploits the operating system's capabilities to gather sensitive information [21, 99]. Here, we newly show that even the seemingly innocuous features on social media could be abused. In the current Twitter design, engagements (following, commenting, etc.) from a protected account send no notifications to the recipients of these actions. This makes both surveillance and surprise harassment campaigns feasible, as discussed earlier.

In a more balanced design, precautions should be taken to avoid possible abuses. For example, private accounts' engagements should generate notifications, or be revealed to the receiver but no one else. In this way, it would be difficult for an attacker to launch surprise harassment campaigns without the target's knowledge. For instance, my harasser would not have been able to secretly follow me and expose others viewing my follower list to doxxing information about me. Alternatively, users could have the option to limit engagement to public accounts only, since if engagement from private accounts is refused, harassment will be much harder. If the content of the account is already fully hidden, then in terms of user privacy, such solutions will not hurt the non-abusive users. It just makes it much harder for the harassers to abuse the privacy features.

6.5 Reporting is Painful

6.5.1 Issue: Reporting Is a Lot of Frustrating Work. While I did not automate reporting initially, I quickly decided it was necessary, because Twitter's reporting process is slow and frustrating.

First, reporting on Twitter involves a lot of cognitive labor, and is very time consuming: one has to click through at least three single/multiple-choice interfaces to tick the right checkboxes before submitting a report. It would take even more time and mental effort (effort that requires me to engage with the harassment), if I choose to provide the optional context information.[10] This is consistent with prior user testimonies [13], where harassment victims receive the least assistance from volunteers despite universally demanding it, due to the challenging nature of reporting.

Second, reporting is emotionally draining and can be traumatizing for the harassment victim. While the emotional labor of both volunteer and commercial moderators has been well documented [119], this labor is not limited to moderators; it also falls on those who initiate the reporting process, typically the victims themselves. During all phases of the reporting process, there are numerous instances where one is forced to read and re-read the content from the harasser, whereas I strongly prefer not to see anything from the harasser at all. For example, if one wants to report individual tweets, one needs to navigate to them and click the report button. When reporting the harasser's profile instead of individual tweets, one is explicitly asked to select the offending tweets, making it unavoidable to read them. This makes it incredibly difficult to avoid reading harmful content, even if one is confident that all tweets from the dedicated harassing account are bad and they would simply like to report blindly.

Third, the responses one gets from reports are unpredictable and can be humiliating. There is no guarantee that the reported accounts will be suspended. For example, my impersonation tickets have never worked despite the fact that the harasser uses my profile

[10] Twitter's reporting flow changed in September 2023. Now the optional context input has been removed. The account here describes what reporting on Twitter was like before September 2023.

image and identity information, and I have submitted my photo ID to prove my identity.

Finally, the feedback given to the user after the reporting is often confusing. You might receive a series of "no violation" emails, but in the end the account is suspended nevertheless. Other times, the account is locked without suspension. In principle, the harasser needs to delete any violating content to unlock the account. However, in my case, since the account name itself is harassing and because my harasser abandons the account soon after being blocked, locking is useless.

6.5.2 Recommendation: Give Targets of Severe Harassment More Privileges in Reporting. While the majority of interactions between abusive accounts and their targets are one-offs, a small minority of users are targets of severe and repeated harassment [80], which can be part of broader violence against the victims [2, 110]. Special protections are needed for targets of severe harassment. For example, my experience has shown that frequent reporting can place a significant mental and emotional burden on the victim. Instead, platforms should simplify the reporting process for users suffering from repeated attacks. Currently, reporting a tweet or a profile requires users to navigate through at least three screens and make selections out of hundreds of possible combinations, which becomes impractical for victims facing high-volume harassment. Although friendsourcing can help alleviate user burdens [86], another approach would provide frequent targets of harassment with additional privileges in the reporting pipeline, streamlining the handling of the reports they originate. For example, when reporting an account, the victim could label it as part of a harassment campaign. The platform could apply the same treatment to that account as it does to others in the campaign, aiding the victim while also reducing the workload for the content moderation team.

While my system incorporates automated reporting, I do so only because I have no viable alternatives. I do not recommend automated reporting as a general solution, given its potential for abuse. When reporting is streamlined for users facing repeated harassment, it is essential to verify that the reported accounts have genuinely engaged in abuse. Otherwise, such privileges could be exploited. For example, a malicious user could coordinate many accounts to fabricate a harassment campaign, obtain elevated reporting privilege, and then misuse it to target other users.

6.5.3 Issue: Reporting Notifications Still Expose Users to Abusive Content despite Blocking. The current design of Twitter reporting system hides the content of the reported tweets in its "We received your report" notification, but leaves the display name and the user handle completely visible. Harassers who know the reporting process can deliberately embed abusive messages in these places, and the victim is forced to read them again and again when checking the progress of reports they have made.

In my case, my harasser routinely uses insulting usernames, which means their efforts successfully face me at every single step before, during, and after the reporting process.

6.5.4 Recommendation: Hiding Abusive Messages More Completely. Some platforms have implemented features like blurring as information hiding mechanisms [27]. In such cases, the platforms leave the option to view more information to the users. However, the platforms need to consider all potential avenues of harassment to apply them effectively. Failing to do so leaves users vulnerable. The display of potentially abusive usernames in Twitter reporting notifications exemplifies such design oversight. In this context, information hiding should be applied universally, with users given the option to reveal hidden information. To help users track their submitted reports, the platform could summarize statuses and outcomes of reports based on submission dates and reasons for reporting, rather than content of accounts or posts.

To arrive at better designs, platforms should create a checklist of locations where users might be exposed to abusive messages at the initial design stage and enlist white hat researchers to identify potential venues for such abuse, as in the case of traditional security research [146]. Twitter, for example, could extend information-hiding features to additional places, such as blurring usernames on the blocked-user list and in updates about submitted reports.

6.6 Recommendation: Providing Special Considerations for Targets of Severe Harassment

As discussed in Section 2.1, the experiences of harassment victims vary widely. Isolated incidents are qualitatively distinct from concerted and targeted campaigns. A one-size-fits-all approach to moderation is unlikely to address the needs of all victims. Nevertheless, I offer several suggestions based on my experience, in addition to what's already discussed in Section 6.5.2.

6.6.1 More Control for Targeted Users. While interaction design has historically sought opportunities for user control [31], scholars have identified downsides to control settings, including the added effort required to find and understand them [122, 133]. However, it would make sense to provide additional options to a select group of highly motivated users, like myself, who are willing to invest more effort in content moderation. For example, the user could have the option to: hide their following/follower list, hide interactions from people with whom they have not interacted, and control post visibility for non-logged-in users. These options would make stalking and harassment significantly harder. Some platforms have implemented a subset of these features [11], but they have not yet been widely adopted.

Platforms could also establish mechanisms that allow targets of severe harassment to access features not available to regular accounts. Alternatively, if made available for all users, "guided tours" have shown potential for revealing useful settings [63]; while most users do not need this, targets of harassment are unusually motivated and might benefit from an approach that showcases all of the options available at different stages of interaction. On a related note, Twitter offers two routes of reporting: via profile/tweets or via Twitter support. Many are unaware of the latter, which offers more options. The platform could streamline and unify these interfaces to increase use. Platforms could also generate visualizations for easier review of all settings.

6.6.2 Different Public Interests Related Design Trade-Offs for Regular Users. It has been argued that content moderation by public figures can have implications for freedom of speech. For instance,

a politician's act of blocking critics on Twitter has been challenged in court, on the grounds that the government cannot exclude individuals solely based on their views [66]. Giving special content moderation treatment to newsworthy content is also controversial [74]. However, here should be a distinction between personal accounts and accounts for public figures or those that serve public interests. Not all design trade-offs made for public interests or newsworthy content need to be applied universally. For instance, on Twitter, the way hiding replies works could be approached differently. Currently, if you hide replies, it substitutes the post with a large banner that actually encourages others to uncover what is hidden there. Not hiding everything completely might make sense for a media/political account, where there may be a public interest in continuing to make those comments available. On the other hand, a victim of targeted harassment — as in my case, where the harasser wanted to make accusations about me to my friends — should have the right to make the unwanted interactions disappear from public view completely.

6.7 Recommendation: Empowering the User with APIs

Finally, platforms need to make APIs readily available to users. Platforms are currently trending in the opposite direction, with increasingly restricted APIs to avoid providing training data to competitors [113]. But as my experience has shown, access to robust APIs is vital for the community to build tools that suit their needs. In fact, the API reads required for building an anti-harassment system as demonstrated in this work is quite limited. The system only needs to read my own posts, interactions directed at me, and information about accounts that have interacted with me. While many individual users may not have the technical skill to leverage these APIs, they can draw on tools built by organizations or third parties that do. And, in contrast to the tools provided by the platform (often minimal and one-size-fits-all), community-built tools are likely address these problems in ways that recognize the lived experience of harassment survivors better. As discussed by Geiger, the notion of "platform sovereignty" is a myth and unilaterally denying users the ability to create much-needed tools for themselves will be counterproductive [42]. My experience demonstrates that it is possible to build highly customizable, automated tools to combat extreme online harassment, and that it is much easier to build them with capable APIs provided by the platform. Commercial incentives should not be allowed to take precedence over allowing users to protect themselves from harm.

7 Conclusion

In response to a prolonged, targeted harassment campaign on Twitter, I developed an automated, personalized, and collaborative anti-harassment system—first using the official Twitter API, then using reverse-engineered Twitter endpoints—to protect myself and my friends. The system successfully reduced my exposure to harassment, resulting in the removal of the majority of the harasser's hundreds of accounts. But this experience demonstrates that even for a technical user with many resources, without good platform-level anti-abuse designs and tech support, it is surprisingly hard to deal with sustained harassment from even a single, committed, non-technical user.

References

[1] Tony E. Adams, Stacy Holman Jones, and Carolyn Ellis. 2021. Introduction: Making sense and taking action: Creating a caring community of autoethnographers. In *Handbook of Autoethnography* (2 ed.), Tony E. Adams, Stacy Holman Jones, and Carolyn Ellis (Eds.). Routledge, New York, NY, USA and Abingdon, UK, 1–19.

[2] Hadeel Al-Alosi. 2017. Cyber-violence : digital abuse in the context of domestic violence. *University of New South Wales Law Journal* 40, 4 (2017), 1573–1603.

[3] Tawfiq Ammari, Sarita Schoenebeck, and Daniel Romero. 2019. Self-declared Throwaway Accounts on Reddit: How Platform Affordances and Shared Norms enable Parenting Disclosure and Support. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 135 (Nov. 2019), 30 pages. doi:10.1145/3359237

[4] Ishaku Hassan Anaobi, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Damilola Ibosiola, and Gareth Tyson. 2023. Will Admins Cope? Decentralized Moderation in the Fediverse. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3109–3120. doi:10.1145/3543507.3583487

[5] Briony Anderson and Mark A. Wood. 2021. Doxxing: A Scoping Review and Typology. In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*. Emerald Publishing Limited, Bingley, UK, 205–226. doi:10.1108/978-1-83982-848-520211015

[6] Carolina Are. 2022. The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies* 22, 8 (2022), 2002–2019. doi:10.1080/14680777.2021.1928259

[7] Carolina Are. 2023. An autoethnography of automated powerlessness: lacking platform affordances in Instagram and TikTok account deletions. *Media, Culture & Society* 45, 4 (2023), 822–840. doi:10.1177/01634437221140531

[8] Tomer Bar. 2018. Notifying Our Developer Ecosystem About a Photo API Bug. https://developers.facebook.com/blog/post/2018/12/14/notifying-our-developer-ecosystem-about-a-photo-api-bug/

[9] Lutfi Basit, Puji Santoso, and Firahmi Rizky. 2025. Multi-platform analysis of sexual harassment networks: gender dynamics and digital amplification. *Social Network Analysis and Mining* 16, 1 (2025), 18. doi:10.1007/s13278-025-01563-3

[10] Heidi R. Biggs, Jeffrey Bardzell, and Shaowen Bardzell. 2021. Watching Myself Watching Birds: Abjection, Ecological Thinking, and Posthuman Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 619, 16 pages. doi:10.1145/3411764.3445329

[11] Bilibili Creator Center. 2023. User Privacy Protection Upgraded! The Settings Now Support "Hiding the Follower List"! Retrieved June 20, 2023 from https://www.bilibili.com/read/cv25076198/

[12] Linda Birt, Suzanne Scott, Debbie Cavers, Christine Campbell, and Fiona Walter. 2016. Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? *Qualitative Health Research* 26, 13 (2016), 1802–1811. doi:10.1177/1049732316654870 PMID: 27340178.

[13] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 24 (dec 2017), 19 pages. doi:10.1145/3134659

[14] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 100 (Nov. 2019), 25 pages. doi:10.1145/3359202

[15] Block Party. 2023. Block Party App. Retrieved June 20, 2023 from https://www.blockpartyapp.com/ archived at [https://web.archive.org/web/20230608024545/https://www.blockpartyapp.com/].

[16] BODYGUARD. 2023. Bodyguard. Retrieved Oct 07, 2023 from https://www.bodyguard.ai/, archived at [https://web.archive.org/web/20230925175002/https://bodyguard.ai/].

[17] Adam Breuer, Roee Eilat, and Udi Weinsberg. 2020. Friend or Faux: Graph-Based Early Detection of Fake Accounts on Social Networks. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1287–1297. doi:10.1145/3366423.3380204

[18] Derek Caelin. 2022. Decentralized Networks vs The Trolls. In *Fundamental Challenges to Global Peace and Security : The Future of Humanity*, Hoda Mahmoudi, Michael H. Allen, and Kate Seaman (Eds.). Springer International Publishing, Cham, 143–168. doi:10.1007/978-3-030-79072-1_8

[19] Jie Cai and Donghee Yvette Wohn. 2019. What are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19 Companion)*. Association for Computing Machinery, New York, NY, USA, 166–170. doi:10.1145/3311957.3359478

[20] Elaine Campbell. 2017. "Apparently Being a Self-Obsessed C**t Is Now Academically Lauded": Experiencing Twitter Trolling of Autoethnographers. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 18, 3 (Sep. 2017), 19 pages. doi:10.17169/fqs-18.3.2819

[21] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. 2018. The Spyware Used in Intimate Partner Violence. In *2018 IEEE Symposium on Security and Privacy (SP)*. 441–458. doi:10.1109/SP.2018.00061

[22] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1217–1230. doi:10.1145/2998181.2998213

[23] Lauren Collins. 2008. Friend Game. The New Yorker. Retrieved October 29, 2024 from https://www.newyorker.com/magazine/2008/01/21/friend-game

[24] Amanda C. Cote. 2017. "I Can Defend Myself": Women's Strategies for Coping With Harassment While Gaming Online. *Games and Culture* 12, 2 (2017), 136–155. doi:10.1177/1555412015587603

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

[26] Sally Jo Cunningham and Matt Jones. 2005. Autoethnography: a tool for practice and education. In *Proceedings of the 6th ACM SIGCHI New Zealand Chapter's International Conference on Computer-Human Interaction: Making CHI Natural* (Auckland, New Zealand) *(CHINZ '05)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/1073943.1073944

[27] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 33–42. doi:10.1609/hcomp.v8i1.7461

[28] Michael Ann DeVito, Ashley Marie Walker, and Jeremy Birnholtz. 2018. 'Too Gay for Facebook': Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 44 (Nov. 2018), 23 pages. doi:10.1145/3274313

[29] 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 (2021), 700–732. doi:10.1007/s12119-020-09790-w

[30] Ángel Díaz and Laura Hecht-Felella. 2021. *Double Standards in Social Media Content Moderation*. Research Report. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

[31] Alan Dix, Janet E. Finlay, Gregory D. Abowd, and Russell Beale. 2003. *Human–Computer Interaction* (3rd ed.). Prentice Hall, Upper Saddle River, NJ, USA.

[32] Harald Dreßing, Josef Bailer, Anne Anders, Henriette Wagner, and Christine Gallas. 2014. Cyberstalking in a Large Sample of Social Network Users: Prevalence, Characteristics, and Impact Upon Victims. *Cyberpsychology, Behavior, and Social Networking* 17, 2 (2014), 61–67. doi:10.1089/cyber.2012.0231

[33] Electronic Frontier Foundation. 2020. Federal Judge Rules It Is Not a Crime to Violate a Website's Terms of Service. Retrieved June 25, 2024 from https://www.eff.org/deeplinks/2020/04/federal-judge-rules-it-not-crime-violate-websites-terms-service

[34] Fatma Elsafoury, Stamos Katsigiannis, Zeeshan Pervez, and Naeem Ramzan. 2021. When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection. *IEEE Access* 9 (2021), 103541–103563. doi:10.1109/ACCESS.2021.3098979

[35] Joseph Farrell and Paul Klemperer. 2007. Chapter 31 Coordination and Lock-In: Competition with Switching Costs and Network Effects. In *Handbook of Industrial Organization*, M. Armstrong and R. Porter (Eds.). Vol. 3. Elsevier, 1967–2072. doi:10.1016/S1573-448X(06)03031-7

[36] Chantal Faucher, Margaret Jackson, and Wandaa Cassidy. 2015. When Online Exchanges Byte: An Examination of the Policy Environment Governing Cyberbullying at the University Level. *Canadian Journal of Higher Education* 45, 1 (2015), 102–121. doi:10.47678/cjhe.v45i1.184215

[37] 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. doi:10.1007/978-3-319-72917-6_16

[38] Brittany Fiore-Silfvast. 2012. User-Generated Warfare: A Case of Converging Wartime Information Networks and Coproductive Regulation on YouTube. *International Journal of Communication* 6 (2012), 1965–1988. https://ijoc.org/index.php/ijoc/article/view/1436

[39] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3174241

[40] Ingo Frommholz, Haider M. al Khateeb, Martin Potthast, Zinnar Ghasem, Mitul Shukla, and Emma Short. 2016. On Textual Analysis and Machine Learning for Cyberstalking Detection. *Datenbank-Spektrum* 16, 2 (2016), 127–135. doi:10.1007/s13222-016-0221-x

[41] Becky Gardiner. 2018. "It's a terrible way to go to work:" what 70 million readers' comments on the Guardian revealed about hostility to women and minorities online. *Feminist Media Studies* 18, 4 (2018), 592–608. doi:10.1080/14680777.2018.1447334

[42] R. Stuart Geiger. 2014. Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society* 17, 3 (2014), 342–356. doi:10.1080/1369118X.2013.873069

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

[44] R. Stuart Geiger and David Ribes. 2010. The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 117–126. doi:10.1145/1718918.1718941

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

[46] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 5 pages. doi:10.1177/2053951720943234

[47] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society* 8, 3 (2022), 20563051221117552. doi:10.1177/20563051221117552

[48] Tarleton Gillespie. 2023. The Fact of Content Moderation; Or, Let's Not Solve the Platforms' Problems for Them. *Media and Communication* 11, 2 (2023), 406–409. doi:10.17645/mac.v11i2.6610

[49] Google Support. 2023. About the YouTube Priority Flagger program. Retrieved June 20, 2023 from https://support.google.com/youtube/answer/7554338?hl=en

[50] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 15 pages. doi:10.1177/2053951719897945

[51] Chandell Gosse, George Veletsianos, Jaigris Hodson, Shandell Houlden, Tonia A. Dousay, Patrick R. Lowenthal, and Nathan Hall. 2021. The hidden costs of connectivity: nature and effects of scholars' online harassment. *Learning, Media and Technology* 46, 3 (2021), 264–280. doi:10.1080/17439884.2021.1878218

[52] 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 242, 17 pages. doi:10.1145/3491102.3517517

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

[54] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 466 (Oct. 2021), 35 pages. doi:10.1145/3479610

[55] Catherine Han, Anne Li, Deepak Kumar, and Zakir Durumeric. 2024. PressProtect: Helping Journalists Navigate Social Media in the Face of Online Harassment. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 509 (Nov. 2024), 34 pages. doi:10.1145/3687048

[56] Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T. Hancock, and Zakir Durumeric. 2023. Hate Raids on Twitch: Echoes of the Past, New Modalities, and Implications for Platform Governance. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 133 (April 2023), 28 pages. doi:10.1145/3579609

[57] Erika Hayasaki. 2023. The Lurker. The Verge. Retrieved October 29, 2024 from https://www.theverge.com/c/features/23903125/lurker-online-harassment-stalking-asian-academics

[58] Daniel Hickey, Matheus Schmitz, Daniel Fessler, Paul E. Smaldino, Goran Muric, and Keith Burghardt. 2023. Auditing Elon Musk's Impact on Hate Speech and Bots. *Proceedings of the International AAAI Conference on Web and Social Media* 17, 1 (Jun. 2023), 1133–1137. doi:10.1609/icwsm.v17i1.22222

[59] Sameer Hinduja and Justin W. Patchin. 2010. Bullying, Cyberbullying, and Suicide. *Archives of Suicide Research* 14, 3 (2010), 206–221. doi:10.1080/13811118.2010.494133 PMID: 20658375.

[60] Jacob Hoffman-Andrews. 2020. BlockTogether. Retrieved November 14, 2023 from https://blocktogether.org/

[61] Sarah Homewood. 2023. Self-Tracking to Do Less: An Autoethnography of Long COVID That Informs the Design of Pacing Technologies. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (, Hamburg, Germany,) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 656, 14 pages. doi:10.1145/3544548.3581505

[62] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv:1702.08138 [cs.LG] https://arxiv.org/abs/1702.08138

[63] Silas Hsu, Kristen Vaccaro, Yin Yue, Aimee Rickman, and Karrie Karahalios. 2020. Awareness, Navigation, and Use of Feed Control Settings Online. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376583

[64] Xiaoyun Huang, Jessica Vitak, and Yla Tausczik. 2020. "You Don't Have To Know My Past": How WeChat Moments Users Manage Their Evolving Self-Presentation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376595

[65] 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. doi:10.1145/3555775

[66] The Knight First Amendment Institute. 2023. Knight Institute v. Trump. Retrieved June 20, 2023 from https://knightcolumbia.org/cases/knight-institute-v-trump

[67] Nazanin Jafari and James Allan. 2025. Reducing the Emotional Distress of Content Moderators through LLM-based Target Substitution in Implicit and Explicit Hate-Speech. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '25)*. Association for Computing Machinery, New York, NY, USA, 255–264. doi:10.1145/3708319.3733712

[68] Sébastien Jambor. 2023. *Understanding ActivityPub - Part 3: The State of Mastodon.* Retrieved January 10, 2024 from https://seb.jambor.dev/posts/understanding-activitypub-part-3-the-state-of-mastodon/ archived at [https://web.archive.org/web/20231221121056/https://seb.jambor.dev/posts/understanding-activitypub-part-3-the-state-of-mastodon/].

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

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

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

[72] Shagun Jhaver and Amy X. Zhang. 2025. Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society* 27, 5 (2025), 2930–2950. doi:10.1177/14614448231217993

[73] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. A Trade-off-centered Framework of Content Moderation. *ACM Trans. Comput.-Hum. Interact.* 30, 1, Article 3 (March 2023), 34 pages. doi:10.1145/3534929

[74] Thomas E. Kadri and Kate Klonick. 2019. Facebook v. Sullivan: Public Figures and Newsworthiness in Online Speech. *Southern California Law Review* 93 (2019), 37–99.

[75] Daphne Keller and Paddy Leerssen. 2020. Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation. In *Social Media and Democracy*, Nathaniel Persily and Joshua A. Tucker (Eds.). Cambridge University Press, 220–251. doi:10.1017/9781108890960

[76] George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology Solutions to Combat Online Harassment. In *Proceedings of the First Workshop on Abusive Language Online*, Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault (Eds.). Association for Computational Linguistics, Vancouver, BC, Canada, 73–77. doi:10.18653/v1/W17-3011

[77] Sarah Khaled, Neamat El-Tazi, and Hoda M. O. Mokhtar. 2018. Detecting Fake Accounts on Social Media. In *2018 IEEE International Conference on Big Data (Big Data)*. 3672–3681. doi:10.1109/BigData.2018.8621913

[78] Sara Kiesler, Robert E. Kraut, Paul Resnick, Aniket Kittur, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. 2011. *Regulating Behavior in Online Communities.* The MIT Press, 125–178. http://www.jstor.org/stable/j.ctt5hhgvw.7

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

[80] Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. 2023. Understanding the Behaviors of Toxic Accounts on Reddit. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 2797–2807. doi:10.1145/3543507.3583522

[81] Lucio La Cava, Sergio Greco, and Andrea Tagarelli. 2021. Understanding the growth of the Fediverse through the lens of Mastodon. *Applied Network Science* 6, 1, Article 64 (2021), 35 pages. doi:10.1007/s41109-021-00392-5

[82] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 54, 18 pages. doi:10.1145/3491102.3501999

[83] Na Yeon Lee and Ahran Park. 2023. How online harassment affects Korean journalists? The effects of online harassment on the journalists' psychological problems and their intention to leave the profession. *Journalism* 25, 4 (2023), 900–920. doi:10.1177/14648849231166511

[84] Ning F. Ma, Veronica A. Rivera, Zheng Yao, and Dongwook Yoon. 2022. "Brush it Off": How Women Workers Manage and Cope with Bias and Harassment in Gender-agnostic Gig Platforms. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 397, 13 pages. doi:10.1145/3491102.3517524

[85] Renkai Ma and Yubo Kou. 2021. "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 429 (Oct. 2021), 25 pages. doi:10.1145/3479573

[86] Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3174160

[87] mastodon. 2019. Hide Replies. Retrieved January 10, 2024 from https://github.com/mastodon/mastodon/issues/11984

[88] mastodon. 2021. [Feature request] Remove replies from threads from user blocked accounts. Retrieved January 10, 2024 from https://github.com/mastodon/mastodon/issues/15631

[89] J. Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, Reviewing, and Responding to Harassment on Twitter. arXiv:1505.03359 [cs.SI] https://arxiv.org/abs/1505.03359

[90] Meta Platforms, Inc. 2023. Instagram Help Center. Retrieved June 20, 2023 from https://help.instagram.com/700284123459336 archived at [https://web.archive.org/web/20230605133112/https://help.instagram.com/700284123459336].

[91] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Tackling Online Abuse: A Survey of Automated Abuse Detection Methods. arXiv:1908.06024 [cs.CL] https://arxiv.org/abs/1908.06024

[92] Moderate. 2023. Moderate App. Retrieved June 20, 2023 from https://moderateapp.com/ archived at [https://web.archive.org/web/20220619214154/https://moderateapp.com/].

[93] Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The Impact of Toxic Language on the Health of Reddit Communities. In *Advances in Artificial Intelligence*, Malek Mouhoub and Philippe Langlais (Eds.). Springer International Publishing, Cham, 51–56. doi:10.1007/978-3-319-57351-9_6

[94] Rachel E. Morgan and Jennifer L. Truman. 2022. *Stalking Victimization, 2019.* Technical Report. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. https://bjs.ojp.gov/library/publications/stalking-victimization-2019

[95] Arvind Narayanan and Sayash Kapoor. 2024. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference.* Princeton University Press, Princeton, NJ, USA.

[96] New York Civil Liberties Union. 2020. What to Do If You're Censored by Politicians on Social Media. NYCLU. Retrieved November 14, 2023 from https://www.nyclu.org/en/know-your-rights/what-do-if-youre-censored-politicians-social-media

[97] Manoj Niverthi, Gaurav Verma, and Srijan Kumar. 2022. Characterizing, Detecting, and Predicting Online Ban Evasion. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2614–2623. doi:10.1145/3485447.3512133

[98] Fayika Farhat Nova, Michael Ann DeVito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. 2021. "Facebook Promotes More Harassment": Social Media Ecosystem, Skill and Marginalized Hijra Identity in Bangladesh. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 157 (April 2021), 35 pages. doi:10.1145/3449231

[99] Christopher Parsons, Adam Molnar, Jakub Dalek, Jeffrey Knockel, Miles Kenyon, Bennett Haselton, Cynthia Khoo, and Ron Deibert. 2019. *The Predator in Your Pocket: A Multidisciplinary Assessment of the Stalkerware Application*

Industry. Citizen Lab Research Report 120. The Citizen Lab, University of Toronto. https://citizenlab.ca/research/the-predator-in-your-pocket-a-multidisciplinary-assessment-of-the-stalkerware-application-industry/

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

[101] Christian Payne. 2002. On the security of open source software. *Information Systems Journal* 12, 1 (2002), 61–78. doi:10.1046/j.1365-2575.2002.00118.x

[102] Filipa Pereira, Brian H. Spitzberg, and Marlene Matos. 2016. Cyber-harassment victimization in Portugal: Prevalence, fear and help-seeking among adolescents. *Computers in Human Behavior* 62 (2016), 136–146. doi:10.1016/j.chb.2016.03.039

[103] Devakunchari Ramalingam and Valliyammai Chinnaiah. 2018. Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers & Electrical Engineering* 65 (2018), 165–177. doi:10.1016/j.compeleceng.2017.05.020

[104] Aravindh Raman, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2019. Challenges in the Decentralised Web: The Mastodon Case. In *Proceedings of the Internet Measurement Conference* (Amsterdam, Netherlands) *(IMC '19)*. Association for Computing Machinery, New York, NY, USA, 217–229. doi:10.1145/3355369.3355572

[105] Amon Rapp. 2018. Autoethnography in Human-Computer Interaction: Theory and Practice. In *New Directions in Third Wave Human-Computer Interaction: Volume 2 - Methodologies*, Michael Filimowicz and Veronika Tzankova (Eds.). Springer International Publishing, Cham, 25–42. doi:10.1007/978-3-319-73374-6_3

[106] Sarah T. Roberts. 2019. *Behind the screen: content moderation in the shadows of social media*. Yale University Press, New Haven, CT, USA.

[107] Pradeep Kumar Roy and Shivam Chahar. 2020. Fake Profile Detection on Social Networking Websites: A Comprehensive Review. *IEEE Transactions on Artificial Intelligence* 1, 3 (2020), 271–285. doi:10.1109/TAI.2021.3064901

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

[109] Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 370 (Nov. 2022), 27 pages. doi:10.1145/3555095

[110] Janneke M. Schokkenbroek, Joris Van Ouytsel, Wim Hardyns, and Koen Ponnet. 2022. Adults' Online and Offline Psychological Intimate Partner Violence Experiences. *Journal of Interpersonal Violence* 37, 15-16 (2022), NP14656–NP14671. doi:10.1177/08862605211015217 PMID: 33966535.

[111] 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 (2019), 1417–1443. doi:10.1177/1461444818821316

[112] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 391, 18 pages. doi:10.1145/3544548.3581538

[113] Umar Shakir. 2023. Reddit's Upcoming API Changes Will Make AI Companies Pony Up. Retrieved June 20, 2023 from https://www.theverge.com/2023/4/18/23688463/reddit-developer-api-terms-change-monetization-ai

[114] Ketaki Shriram and Raz Schwartz. 2017. All are welcome: Using VR ethnography to explore harassment behavior in immersive social virtual reality. In *2017 IEEE Virtual Reality (VR)*. 225–226. doi:10.1109/VR.2017.7892258

[115] Guy Simon. 2022. OpenWeb tests the impact of "nudges" in online discussions. Retrieved October 20, 2024 from https://www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api

[116] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 107, 20 pages. doi:10.1145/3544548.3581057

[117] Andrew C. Sparkes. 2021. When Judgment Calls: Making Sense of Criteria for Evaluating Different Forms of Autoethnography. In *Handbook of Autoethnography* (2 ed.), Tony E. Adams, Stacy Holman Jones, and Carolyn Ellis (Eds.). Routledge, 263–276. doi:10.4324/9780429431760

[118] Nick Srnicek. 2017. *Platform Capitalism*. Polity Press, Cambridge, UK.

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

[120] 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 (2021), 367–376. doi:10.1089/cyber.2020.0253 PMID: 33181026.

[121] Angelika Strohmayer, Julia Slupska, Rosanna Bellini, Lynne Coventry, Tara Hairston, and Adam Dodge. 2021. *Trust and Abusability Toolkit: Centering Safety in Human-Data Interactions*. Research Report. Human-Data Interaction Network+ (UKRI). https://researchportal.northumbria.ac.uk/en/publications/trust-and-abusability-toolkit-centering-safety-in-human-data-inte/

[122] Hyewon Suh, Nina Shahriaree, Eric B. Hekler, and Julie A. Kientz. 2016. Developing and Validating the User Burden Scale: A Tool for Assessing User Burden in Computing Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3988–3999. doi:10.1145/2858036.2858448

[123] Nicolas Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Towards Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13 (2019), 1526–1543. https://ijoc.org/index.php/ijoc/article/view/9736

[124] Kejsi Take, Victoria Zhong, Chris Geeng, Emmi Bevensee, Damon McCoy, and Rachel Greenstadt. 2024. Stoking the Flames: Understanding Escalation in an Online Harassment Community. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 176 (April 2024), 23 pages. doi:10.1145/3641015

[125] 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)*. 247–267. doi:10.1109/SP40001.2021.00028

[126] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. 2022. "It's common and a part of being a content creator": Understanding How Creators Experience and Cope with Hate and Harassment Online. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 121, 15 pages. doi:10.1145/3491102.3501879

[127] Jay Daniel Thompson. 2023. Beginning of the end: how Elon Musk's removal of the block function on X could trigger its hellish demise. Retrieved December 31, 2023 from https://theconversation.com/beginning-of-the-end-how-elon-musks-removal-of-the-block-function-on-x-could-trigger-its-hellish-demise-211897

[128] Michail Tsikerdekis and Sherali Zeadally. 2014. Multiple Account Identity Deception Detection in Social Media Using Nonverbal Behavior. *IEEE Transactions on Information Forensics and Security* 9, 8 (2014), 1311–1321. doi:10.1109/TIFS.2014.2332820

[129] Sarah Turner, Jason R.C. Nurse, and Shujun Li. 2022. "It Was Hard to Find the Words": Using an Autoethnographic Diary Study to Understand the Difficulties of Smart Home Cyber Security Practices. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 34, 8 pages. doi:10.1145/3491101.3503577

[130] Twitter. 2023. Advanced Twitter Mute Options. Retrieved 2023-06-20 from https://help.twitter.com/en/using-twitter/advanced-twitter-mute-options archived at [https://web.archive.org/web/20230620224424/https://help.twitter.com/en/using-twitter/advanced-twitter-mute-options].

[131] Twitter. 2023. Twitter API. Retrieved June 20, 2023 from https://developer.twitter.com/en/products/twitter-api archived at [https://web.archive.org/web/20230705150448/https://developer.twitter.com/en/products/twitter-api].

[132] U.S. Attorney's Office, Eastern District of New York. 2023. Police Charged with Perpetrating Transnational Repression Scheme Targeting U.S. Residents. Retrieved December 31, 2023 from https://www.justice.gov/d9/2023-04/squad_912_-_23-mj-0334_redacted_complaint_signed.pdf

[133] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173590

[134] Kristen Vaccaro, Karrie Karahalios, Christian Sandvig, Kevin Hamilton, and Cedric Langbort. 2015. Agree or cancel? Research and terms of service compliance. In *ACM CSCW Ethics Workshop: Ethics for Studying Sociotechnical Systems in a Big Data World*. 6 pages.

[135] Noelia Valenzuela-García, Diego J. Maldonado-Guzmán, Andrea García-Pérez, and Cristina Del-Real. 2023. Too Lucky to Be a Victim? An Exploratory Study of Online Harassment and Hate Messages Faced by Social Media Influencers. *European Journal on Criminal Policy and Research* 29, 3 (2023), 397–421. doi:10.1007/s10610-023-09542-0

[136] The Verge. 2023. Anti-harassment service Block Party leaves Twitter amid API changes. Retrieved June 20, 2023 from https://www.theverge.com/2023/5/31/23743538/block-party-hiatus-twitter-app-anti-harassment-service-api
[137] Janet Vertesi. 2022. Op-Ed: I hid my pregnancy from the internet so I know: Online privacy is nearly impossible. Retrieved October 20, 2024 from https://www.latimes.com/opinion/story/2022-05-16/pregnancy-internet-online-privacy-impossible
[138] Jessica Vitak, Stacy Blasiola, Sameer Patil, and Eden Litt. 2015. Balancing Audience and Privacy Tensions on Social Network Sites: Strategies of Highly Engaged Users. *International Journal of Communication* 9 (2015), 1485–1504. https://ijoc.org/index.php/ijoc/article/view/3208
[139] 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* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1231–1245. doi:10.1145/2998181.2998337
[140] Emily A. Vogels. 2021. The State of Online Harassment. Retrieved June 20, 2023 from https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/
[141] Emma Vossen. 2018. *On the Cultural Inaccessibility of Gaming: Invading, Creating, and Reclaiming the Cultural Clubhouse.* PhD dissertation. University of Waterloo.
[142] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Franziska Roesner, and Kurt Thomas. 2023. "There's so much responsibility on users right now:" Expert Advice for Staying Safer From Hate and Harassment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 190, 17 pages. doi:10.1145/3544548.3581229
[143] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2023. Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. *ACM Trans. Comput.-Hum. Interact.* 30, 6, Article 83 (Sept. 2023), 36 pages. doi:10.1145/3603625
[144] Jing Zeng and D. Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet* 14, 1 (2022), 79–95. doi:10.1002/poi3.287
[145] Andy Zhao and Zhaodi Chen. 2023. Let's report our rivals: how Chinese fandoms game content moderation to restrain opposing voices. *Journal of Quantitative Description: Digital Media* 3 (2023), 35 pages. doi:10.51685/jqd.2023.006
[146] Mingyi Zhao, Jens Grossklags, and Peng Liu. 2015. An Empirical Study of Web Vulnerability Discovery Ecosystems. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery, New York, NY, USA, 1105–1117. doi:10.1145/2810103.2813704

A Twitter vs. Other Platforms

Through a series of experiments with multiple accounts, I have systematically tested the functionality offered for common content moderation features (e.g., muting, blocking, hiding) on different platforms, to understand how representative Twitter's implementations are of social media in general. The results of this are shared in Table 1. I also use this to provide context for how my ideal version of blocking and restricting would function, noting that Instagram blocks are very near my optimal solution, but no platform offers restriction functionality that addresses my needs.

Visibility Control. While Twitter is bad in giving controls to users with engagements they receive, it is not alone. There is no way to hide or remove unwanted engagements on Mastodon [87] without contacting the admin of the harasser's home instance. Allowing the user to delete comments could bring about its own problem. On Youtube, Instagram, Facebook and Sina Weibo, anyone can delete comments made to their posts. This has raised questions about censorship, and civil liberty organizations have even published tips for users when a public official deletes their comments unfairly [96]. As discussed in Section 6.6.2, different treatments of accounts of public interests and accounts from private citizens are also needed.

Moderation Operations Other than Mutes and Blocks. Blocking that only affects the visibility of the content between the blocker and the blockee is not uncommon. Mastodon has a similar design [88]. On the other hand, there are platforms that implement user-initiated "shadowban" functionalities that hide the interaction from other users without notifying them. On Youtube, the "hide user from channel" feature allows a channel to make all comments from the harasser invisible. On Instagram, when you block someone, it automatically removes all engagement from the blocked account on your posts, making them unavailable to third parties. Additionally, you have the option to restrict users, requiring their engagements to gain your approval before being visible to anyone other than themselves. For a victim of targeted harassment, Youtube-style hide and Instagram-style restriction might be preferable over blocking and muting, as they allow the victim to control the visibility of the harassing comments without immediately alerting the harasser.

Implementation of Blocking. Twitter is not the only platform where asymmetrically implemented blocking enables abuse: based on my own experience, Sina Weibo suffers from the same problem.

	Interactivity		Visibility				Scope			Alert
If I do this to others	Am I barred from interacting with them?	Are they barred from interacting with me?	Are my posts invisible to them?	Are their interactions with me invisible to me?	Are my interactions with them invisible to 3rd party (guest included)?	Are their interactions with me invisible to 3rd party (guest included)?	Retrospectively applied?	Visibility changes fully reversible?	Applied to other accounts created by them?	Do they know about it immediately?
Mutes	○	○	○	○	○	○	○	○	○	○
Twitter block	⊛	⊛	●	●	○	○	●	●	○	●
Instagram/Threads block	●	●	●	●	⊛	●	●	⊛	●	●
Mastodon block	●	●	●	●	○	○	●	●	○	●
Sina Weibo block	○	●	●	●	○	○	●	●	○	●
Instagram/Threads restriction	○	○	○	●	○	●	○	●	○	○
Youtube hide	⊛	○	○	○	○	●	●	⊛	○	○
My ideal (block-like)	●	●	●	●	●	●	●	●	●	●
My ideal (restriction-like)	●	○	○	●	●	●	●	●	●	○

Table 1: Properties of account-level moderation actions on platforms that I regularly use, and my preferred combinations that would make managing a harassment campaign easier. A ● indicates "Yes", a ○ indicates "No", and a ⊛ indicates "True for some, but not all situations". "Guest" refers to a user who is not logged-in.