Digital Harassment

Treating Online Abuse Like Spam

BY
YAEL GRAUER, Program Manager, Cybersecurity Research, Consumer Reports
DEEPAK KUMAR, Assistant Professor, Computer Science & Engineering, UC San Diego
VIKTORYA VILK, Director, Digital Safety and Free Expression, PEN America

MAY 21, 2025

PEN AMERICA | CR Consumer Reports®

Contents

Summary

Online abuse is a widespread and very real problem. According to a 2021 study by the Pew Research Center, nearly half of Americans have experienced online harassment,[1] which can damage mental health, cause self-censorship, and even put lives at risk. For public-facing professionals—such as journalists, scholars, politicians, and creators—the rates of exposure are even higher. Yet technology companies, including social media platforms, are not sufficiently investing in mechanisms that alleviate the harms of online hate and harassment. They remain overly reliant on reactive solutions, which require users to encounter online abuse–often repeatedly–before it can be addressed, rather than investing in solutions that proactively reduce exposure to abuse while protecting free expression.

In this report, we set out to explore an innovative idea that first emerged in 2018: What would happen if technology companies treated online abuse more like spam? In other words, what if platforms empowered individual users with a mechanism that automatically detected potentially abusive content proactively and quarantined it, so that users could then choose to review and address it—or ignore it altogether? We examine the pros, cons, and nuances involved in this proposal, from both a technical and sociocultural standpoint. We conclude with a detailed set of recommendations outlining how technology companies can implement proactive abuse mitigation mechanisms like the one proposed here. We intend this report to guide technologists, policy experts, trust and safety experts, and researchers who are committed to reducing the negative impacts of online abuse in order to make online spaces safer, more equitable, and more free.

Introduction

Online Abuse Chills Free Expression

Online abuse is making public discourse, so much of which now plays out in digital spaces, less equitable and less free. Malicious actors, ranging from disaffected trolls to state-sponsored cyber-armies, deploy violent threats, hateful slurs, sexual harassment, doxing, and other abusive tactics to intimidate, discredit, and silence their targets.[2]

These tactics are pervasive. According to a 2021 study from the Pew Research Center, 41% of regular users in the U.S. have experienced online harassment, and the volume and severity of such attacks are increasing. People are often targeted because of their profession, identity, and/or political beliefs.[3] Online abuse disproportionately impacts those whose voices have been

[1] Emily Vogels, "The State of Online Harassment," *Pew Research Center*, January 13, 2021, pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/; see also Kurt Thomas et al., "SoK: Hate, Harassment, and the Changing Landscape of Online Abuse," *2021 IEEE Symposium on Security and Privacy,* May 2021, 247-267, ieeexplore.ieee.org/document/9519435.

[2] In this report, we use the terms "online abuse" and "online harassment" interchangeably; PEN America defines these terms as the "pervasive or severe targeting of an individual or group online through harmful behavior." "Defining 'Online Abuse': A Glossary of Terms," Online Harassment Field Manual, *PEN America,* onlineharassmentfieldmanual.pen.org/defining-online-harassment-a-glossary-of-terms/.

[3] Emily Vogels, "The State of Online Harassment," *Pew Research Center*, January 13, 2021, pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/; see also Kurt Thomas et al., "SoK: Hate, Harassment, and

historically marginalized: women,[4] people of color,[5] LGBTQ+ individuals,[6] and members of religious or ethnic minorities.[7] It is particularly pernicious for people who need to have an online presence to do their jobs, including journalists,[8] academics and researchers,[9] and content creators.[10]

Online harassment is an especially prevalent problem for women journalists—a 2022 study from UNESCO and the International Center for Journalists (ICFJ) found that 73% of women respondents had experienced it. Such tactics can take a serious toll on mental health, and, in some cases, even migrate offline. As part of the aforementioned study, 20% of women journalists reported that they had been attacked offline in incidents connected to online abuse. Given the severity and level of risk, online harassment can be extremely effective in stifling free expression, with targeted individuals censoring themselves, leaving online platforms, and sometimes even leaving their professions altogether.[11]

---

the Changing Landscape of Online Abuse," 2021 IEEE Symposium on Security and Privacy, May 2021, 247-267, ieeexplore.ieee.org/document/9519435.

[4] Marjan Nadim and Audun Fladmoe, "Silencing Women? Gender and Online Harassment," *Social Science Computer Review* 39, no. 2 (July 30, 2019), 245-258, journals.sagepub.com/doi/10.1177/0894439319865518; see also Emily Vogels, "The State of Online Harassment," *Pew Research Center,* January 13, 2021, pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

[5] Maeve Duggan, "1 in 4 Black Americans Have Faced Online Harassment Because of Their Race or Ethnicity," *Pew Research Center*, July 25, 2017, pewresearch.org/short-reads/2017/07/25/1-in-4-black-americans-have-faced-online-harassment-because-of-their-race-or-ethnicity/; see also Emily Vogels, "The State of Online Harassment," *Pew Research Center,* January 13, 2021, pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

[6]Abreu, R.L., & Kenny, M.C., "Cyberbullying and LGBTQ Youth: A Systematic Literature Review and Recommendations for Prevention and Intervention," *Journal of Child & Adolescent Trauma*, 11(1) (2017), 81-97. doi.org/10.1007/s40653-017-0175-7.

[7] Niloufar Salehi et al., "Sustained Harm Over Time and Space Limits the External Function of Online Counterpublics for American Muslims," *ACM on Human-Computer Interaction* 7, no. 93 (April 2023), 1-24, dl.acm.org/doi/abs/10.1145/3579526 ; Brooke Auxier, "About One-in-five Americans Who Have Been Harassed Online Say It Was Because of Their Religion," *Pew Research Center*, February 1, 2021, pewresearch.org/short-reads/2021/02/01/about-one-in-five-americans-who-have-been-harassed-online-say-it-was-because-of-their-religion/.

[8] Ferrier, M.P.B. (2018). "Attacks and Harassment: The Impact on Female Journalists and their Reporting," *TrollBusters and the International Women's Media Foundation,* https://www.iwmf.org/attacks-and-harassment/; Lucy Westcott, "'The threats follow us home': Survey details risks for female journalists in U.S., Canada," *Committee to Protect Journalists,* September 4, 2019, cpj.org/2019/09/canada-usa-female-journalist-safety-online-harassment-survey/; Julie Posetti and Nabeelah Shabbir, "The Chilling: A Global Study On Online Violence Against Women Journalists," *International Center for Journalists,* November 2, 2022, icfj.org/our-work/chilling-global-study-online-violence-against-women-journalists; Michelle Ferrier and Nisha Garud-Patkar, "TrollBusters: Fighting Online Harassment of Women Journalists," *Mediating Misogyny* (2018): 311-332. https://doi.org/10.1007/978-3-319-72917-6_16.

[9] Atte Oksanen, Magdalena Celuch, Rita Latikka, Reetta Oksa, and Nina Savela, "Hate and Harassment in Academia: The Rising Concern of the Online Environment," *Higher Education* 84 (November 23, 2021): 541–567, doi.org/10.1007/s10734-021-00787-4; Naomi Nix, Joseph Menn, "These Academics Studied Falsehoods Spread by Trump. Now the GOP Wants Answers," *The Washington Post*, June 6, 2023, washingtonpost.com/technology/2023/06/06/disinformation-researchers-congress-jim-jordan/; Bianca Nogrady, "'I Hope You Die': How the COVID Pandemic Unleashed Attacks on Scientists," *Nature,* October 13, 2021, nature.com/articles/d41586-021-02741-x.

[10] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein, "'It's Common and a Part of Being a Content Creator': Understanding How Creators Experience and Cope with Hate and Harassment Online," *CHI Conference on Human Factors in Computing Systems*, no. 121 (April 27, 2022), 1-15, dl.acm.org/doi/fullHtml/10.1145/3491102.3501879.

[11] Julie Posetti and Nabeelah Shabbir, "The Chilling: A Global Study On Online Violence Against Women Journalists," *International Center for Journalists,* November 2, 2022, icfj.org/our-work/chilling-global-study-online-violence-against-women-journalists; "Online Harassment Survey: Key Findings," *PEN America,* 2017, pen.org/online-harassment-survey-key-findings/; Michelle Ferrier and Nisha Garud-Patkar, "TrollBusters:

Technology Companies Are Backsliding

Social media platforms—such as Facebook, Instagram, TikTok, X (formerly Twitter), and YouTube, where so much of the abuse unfolds—are not doing enough to protect and support their most vulnerable users. In the aforementioned 2021 Pew Research Center study, more than half of Americans saw online harassment as a major problem, yet only 18% felt that social media platforms were doing a good or excellent job of addressing it.[12]

There is a great deal that platforms can do to mitigate online abuse while protecting free expression. They can strengthen harassment, hate speech, and bullying policies by keeping up with ever-evolving tactics.[13] They can improve the implementation of their policies by bolstering behind-the-scenes content moderation by both human moderators and automated systems, including by creating more flexible and effective harassment reporting processes.[14] They can create proactive tools that shield users from the worst abuse, improve reactive tools that allow users to block, mute, and otherwise take action on abusive content, and introduce more effective accountability tools, such as escalating penalties for repeated policy violations and nudges that prompt users to rethink potentially abusive content before they post.[15]

Over the past decade, many platforms had actually started to reform their policies and add new product features to reduce abuse, albeit only after significant advocacy from affected users and civil society (non-governmental organizations, community groups, and special interest groups). To give just one example, Twitter (now X) only introduced an in-app "report abuse" function in 2013, seven years after the company launched and only after public outcry, including a petition that garnered over a hundred thousand signatures to protest coordinated online attacks against prominent women.[16]

Industry incentives for effectively addressing online abuse are mixed. On the one hand, government initiatives to regulate social media platforms are on the rise globally. The EU Digital Services Act (DSA), for example, which was passed in 2022, now requires social media platforms and search engines to "identify, analyse, and assess systemic risks that are linked to their services."[17] The U.K. Online Safety Act, passed in 2023, establishes a duty of care for

---

Fighting Online Harassment of Women Journalists," *Mediating Misogyny* (2018): 311-332. https://doi.org/10.1007/978-3-319-72917-6_16.

[12] Emily A. Vogel, "The State of Online Harassment," *Pew Research Center,* January 13, 2021, pewresearch.org/internet/2021/01/13/americans-views-on-how-online-harassment-should-be-addressed/.

[13] Platforms use a variety of different terms in their policies—primarily "harassment," "cyberbullying/ bullying," and "hateful conduct/hate speech." In this report, we use the terms "online abuse" and "online harassment," except in cases where we are discussing a platform's specific policy, in which case we use the terminology used by that platform in that policy.

[14] "Shouting into the Void: Why Reporting Abuse to Social Media Platforms is So Hard and How to Fix It," *PEN America,* June 29, 2023, pen.org/report/shouting-into-the-void/.

[15] **"**No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users," *PEN America,* March 2021, pen.org/report/no-excuse-for-abuse/.

[16] "Twitter Unveils New Tools to Fight Harassment," *CBS News,* March 28, 2021, cbsnews.com/video/twitter-unveils-new-tools-to-fight-harassment/; Keith Moore, "Twitter 'Report Abuse' Button Calls After Rape Threats," *BBC News,* July 27, 2013, bbc.com/news/technology-23477130; Alexander Abad-Santos, "Twitter's 'Report Abuse' Button Is a Good, But Small, First Step," *The Atlantic*, July 31, 2013, theatlantic.com/technology/archive/2013/07/why-twitters-report-abuse-button-good-tiny-first-step/312689/.

[17] "DSA: Very Large Online Platforms and Search Engines," *European Commission,* August 25, 2023, digital-strategy.ec.europa.eu/en/policies/dsa-vlops#:~:text=This%20means%20that%20they%20must,consumer%20protection%20and%20children's%20rights.

online platforms to protect their users from harmful content.[18] At the same time, because sustaining user attention and maximizing engagement underpins the business model of most social media companies, they build their platforms to prioritize immediacy, emotional impact, and virality, which often serves to amplify abusive behavior.[19]

Unfortunately, over the past two years, technology companies have been backsliding, making their priorities abundantly clear. In 2023, among industry-wide layoffs of nearly 200,000 tech-sector employees, platforms significantly or entirely slashed their trust and safety teams, which are primarily responsible for combating online abuse.[20] Some platforms have even rolled back, or threatened to remove, vital anti-harassment mechanisms. X, for example, rolled back its deadnaming policy and weakened its block button.[21] Meta recently loosened its hate speech policies to allow more abusive content targeting women, LGBTQ+ individuals, and immigrants and significantly scaled back its platform-wide automated moderation systems in favor of user reporting. Experts predict that these changes will increase the proliferation of hate and abuse on Meta's platforms and put an even greater onus on its users to navigate such content on their own.[22]

Proactive vs. Reactive Mechanisms for Addressing Abuse

Today, platforms are overreliant on features that address online abuse *reactively,* such as reporting and blocking, which require users to encounter abuse–often repeatedly–in order to mitigate it. The problem is that frequent exposure to online harassment can have serious consequences. A study by UNESCO found that online abuse negatively impacted the mental health of its targets, with 26% of subjects reporting depression, anxiety, PTSD, and other stress-related ailments like sleep loss and chronic pain.[23] In extreme cases, targeted individuals have contemplated and even died by suicide.[24] The psychological toll has also been shown to

[18] "Online Safety Act: Explainer," *U.K. Department for Science, Innovation & Technology,* May 8, 2024, gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer.
[19] Amit Goldenberg and James J. Gross, "Digital Emotion Contagion," *Harvard Business School,* 2020, hbs.edu/faculty/Publication%20Files/digital_emotion_contagion_8f38bccf-c655-4f3b-a66d-0ac8c09adb2d.pdf; Luke Munn, "Angry By Design: Toxic Communication and Technical Architectures," *Humanities and Social Sciences Communications* 7, no. 53 (2020), doi.org/10.1057/s41599-020-00550-7; Molly Crockett, "How Social Media Amplifies Moral Outrage," *The Eudemonic Project,* February 9, 2020, eudemonicproject.org/ideas/how-social-media-amplifies-moral-outrage.
[20] "The Crunchbase Tech Layoffs Tracker," *Crunchbase News,* news.crunchbase.com/startups/tech-layoffs/; Vittoria Elliot, "Big Tech Ditched Trust and Safety. Now Startups Are Selling It Back As a Service," *WIRED*, November 6, 2023, wired.com/story/trust-and-safety-startups-big-tech/.
[21] Emma Roth and Kylie Robison, "X will let people you've blocked see your posts," *The Verge,* September 23, 2024, theverge.com/2024/9/23/24252438/x-blocked-users-view-public-posts; Nora Benavidez, "Big Tech Backslide: How Social Media Rollbacks Endanger Democracy Ahead of the 2024 Elections," *Free Press,* December 2023, freepress.net/big-tech-backslide-report; Leanna Garfield and Jenni Olson. "All Social Media Platform Policies Should Recognize Targeted Misgendering and Deadnaming as Hate Speech," *GLAAD,* March 5, 2024. glaad.org/social-media-platform-policies-targeted-misgendering-deadnaming-hate-speech/.
[22] Sarah Gilbert, "Three reasons Meta will struggle with community fact-checking," *MIT Technology Review,* January 29, 2025, technologyreview.com/2025/01/29/1110630/three-reasons-meta-will-struggle-with-community-fact-checking/; Justine Calma, "Meta is leaving its users to wade through hate and disinformation," *The Verge,* January 7, 2025, theverge.com/2025/1/7/24338127/meta-end-fact-checking-misinformation-zuckerberg.
[23] Julie Posetti, Nabeelah Shabbir et al., "The Chilling: Global Trends in Online Violence Against Women Journalists," *UNESCO,* April 2021, 12, unesdoc.unesco.org/ark:/48223/pf0000377223.
[24] Jackson Richman, "Taylor Lorenz Breaks Down on MSNBC Sharing Experience Being Targeted Online, Contemplated Suicide," *Mediaite*, April 1, 2022, mediaite.com/tv/taylor-lorenz-breaks-down-on-msnbc-sharing-experience-being-targeted-online-contemplated-suicide/; Ben Dooley and Hikari Hida, "After Reality Star's Death, Japan Vows to Rip the Mask Off Online Hate," *The New York Times*, June 1, 2020, nytimes.com/2020/06/01/business/hana-kimura-terrace-house.html.

cause some public-facing professionals, like journalists, to leave their jobs.[25] It is imperative that technology companies invest in measures that not only address abuse *reactively* but also reduce users' exposure to abuse *proactively*.

Treating Online Abuse More Like Spam?

One innovative proposal to proactively reduce exposure to online abuse—explored by technology journalist Sarah Jeong in her 2015 book "The Internet of Garbage"—is to treat it more like spam.[26] This idea is inspired by email filters, which have largely been successful in relegating spam to separate folders that users can choose whether or not to interact with. Free expression nonprofit PEN America resurfaced Jeong's idea in its 2021 report "No Excuse for Abuse," proposing that platforms give individual users access to a sophisticated mechanism that proactively detects potentially abusive content and quarantines in a centralized area, where users could review the content and decide how to address it—or ignore it altogether.[27] In order to operate with some degree of automation, such a system would need to be powered, at least in part, by artificial intelligence (AI) and/or large language models (LLMs).[28]

Most major social media platforms already rely on a combination of automation and human moderation behind the scenes to proactively identify certain kinds of harmful content in order to reduce its reach, label it, hide it behind screens, or delete it altogether—for all users.[29] Detecting and adjudicating potentially abusive content at a global scale across hundreds of languages and sociopolitical contexts is inherently challenging and complex. An enormous amount of potentially abusive content is perceptual, contextual, and falls into a gray area, with users disagreeing about whether certain content crosses the line into abuse according to their own individual experience and understanding. Platform-driven moderation systems can and do make mistakes, sometimes failing to protect users from abuse while simultaneously suppressing free expression.

One of the primary benefits of the aforementioned innovative proposal is that it gives individual users access to some of the powerful tools that platforms are already using, thereby giving users more control over the content they see and interact with. Some users may not want to see any potentially abusive content at all, while others may, for example, need to sift through banal personal attacks in order to track credible death threats. The point, however, is that users themselves should be able to exercise significantly more agency over their own social media experience.

---

[25] Katherine Goldstein, "When Harassment Drives Women out of Journalism," *International Women's Media Foundation,* 2017, iwmf.org/2017/12/when-harassment-drives-women-out-of-journalism/.
[26] Annalee Newitz, "What if We Treated Online Harassment the Same Way We Treat Spam?" *Ars Technica,* June 23, 2016, arstechnica.com/tech-policy/2016/06/what-if-we-treated-online-harassment-the-same-way-we-treat-spam/; Sarah Jeong, "The Internet of Garbage," *Forbes,* 2015, republished on *Vox,* 2018, cdn.vox-cdn.com/uploads/chorus_asset/file/12599893/The_Internet_of_Garbage.0.pdf.
[27] Viktorya Vilk, Elodie Vialle, and Matt Bailey, "No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users," *PEN America,* March 31, 2021, pen.org/report/no-excuse-for-abuse/.
[28] Shagun Jhaver et al., "Designing Word Filter Tools for Creator-led Comment Moderation," *2022 CHI Conference on Human Factors in Computing Systems*, no. 205 (April 2022), 1-21, dl.acm.org/doi/abs/10.1145/3491102.3517505; Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric, "Watch Your Language: Investigating Content Moderation with Large Language Models," *Human-Computer Interaction (cs.HC),* January 2024, arxiv.org/abs/2309.14517.
[29] Viktorya Vilk, Elodie Vialle, and Matt Bailey, "No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users," *PEN America,* March 31, 2021, pen.org/report/no-excuse-for-abuse/; email to PEN America from Reddit spokesperson, March 2025; email to PEN America from Meta spokesperson, January 2023.

What You Will Find in This Report

In this report, PEN America has joined forces with Consumer Reports, the nonprofit research, testing and advocacy organization, to explore in greater depth what would happen if technology companies treated online abuse more like spam. Like Jeong, we use the comparison with spam not literally, but as metaphor for the idea of platforms giving individual users more robust automated mechanisms to proactively detect potentially abusive content and quarantine it, so that users could choose whether to review, address it, or ignore it. In Section I, we map out how a mechanism like this could work. In Sections II and III, we explore whether the technology to automatically detect abusive content currently exists, analyzing in-platform and third-party tools that already perform some of these functions and examining how this landscape is evolving with the advent of LLMs and generative AI (GenAI) technologies. In Section IV, we examine why users do not already have access to more robust mechanisms to proactively detect and isolate abusive content, including platform incentive structures and priorities. In Section V, we discuss the challenges of proactively detecting and quarantining abusive content, from implicit bias to explicit censorship, and map out ways these challenges can be addressed. In Sections VI and VII, we end with an exploration of the advantages of treating online abuse more like spam and provide a detailed set of recommendations outlining how technology companies can put such proactive anti-abuse measures into practice.

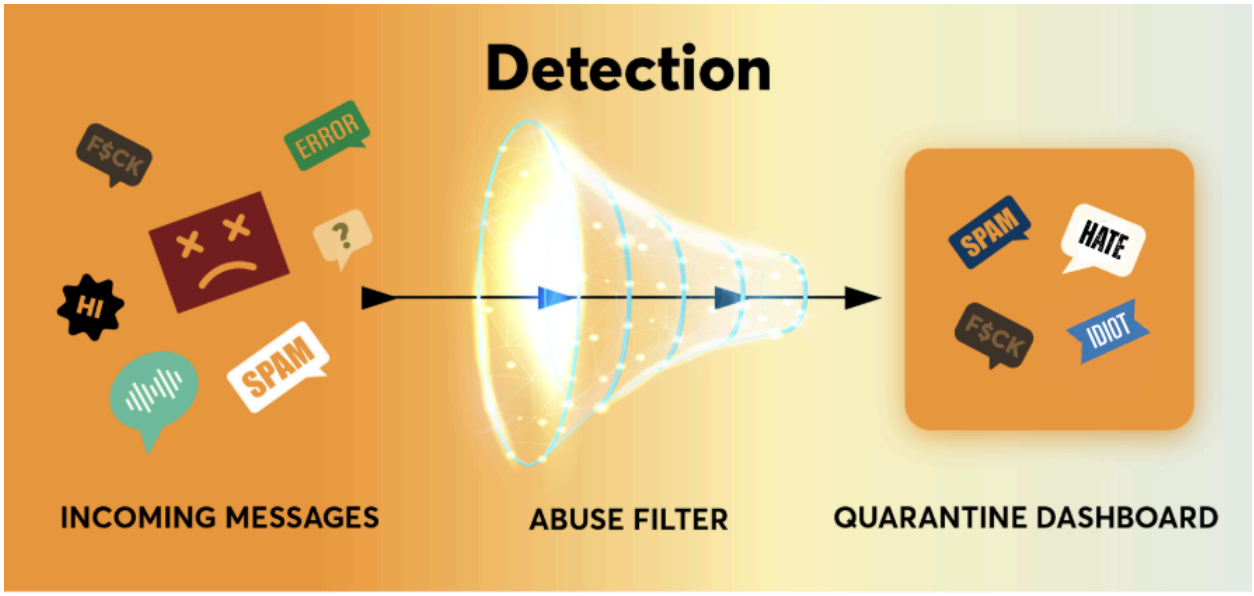
We intend this report to guide technologists, policy experts, trust and safety experts, and researchers in the technology, civil society, and academic sectors who are looking for solutions to better protect and support people facing online abuse. It is important to note that we intend this proposal to be one intervention among many that platforms deploy to reduce online harassment and hate. A user-driven mechanism like the one that we propose here cannot—and should not—replace platform-driven content moderation. This would unduly place the onus on individual users, instead of on well-resourced and powerful platforms and their own internal content moderation teams. We need both individualized tools that allow users to shape their online experiences and rigorous platform-driven content moderation efforts to ensure that online spaces are truly safe, equitable, and free.

Section I: Taking Inspiration From Spam to Reduce Exposure to Online Abuse

In the early days of email in the 1990s, people were bombarded with spam. Within a decade, all major email providers had integrated spam filters that automatically identified junk mail and isolated it in designated spam folders to keep it from cluttering inboxes.[30] With the advent of social media, spam detection was also integrated into platforms to reduce the visibility of junk in feeds, direct messages (DMs), etc. Although there are key differences that make filtering online abuse significantly more difficult than filtering spam—which we explore in detail in Section V—lessons could be learned from the world of spam, where automated filtering has largely been a success.

[30] Lindsay Tjepkama, "A (Brief) History of Spam Filtering and Deliverability," *Emarsys,* December 19, 2013, emarsys.com/learn/blog/a-brief-history-of-spam-filtering-and-deliverability-gunter-haselberger/#:~:text=The%20late%201990s%20%2F%20early%202000s,keywords%2C%20patterns%20or%20special%20characters.

In its 2021 report "No Excuse for Abuse," PEN America—building on an idea sketched out in technology journalist Sarah Jeong's 2018 book "The Internet of Garbage"—recommended a mechanism that would "proactively filter abusive content (across feeds, threads, comments, replies, direct messages, etc.) and quarantine it in a dashboard, where [users could] review and address [the content] as needed with the help of trusted allies."[31] In other words, technology companies could take inspiration from email spam folders to reduce individual users' exposure to hate and harassment on social media platforms, in email, and beyond. Below we outline how such a mechanism could work in three phases—**detection, review,** and **action**—and offer more detailed recommendations in Section VII.

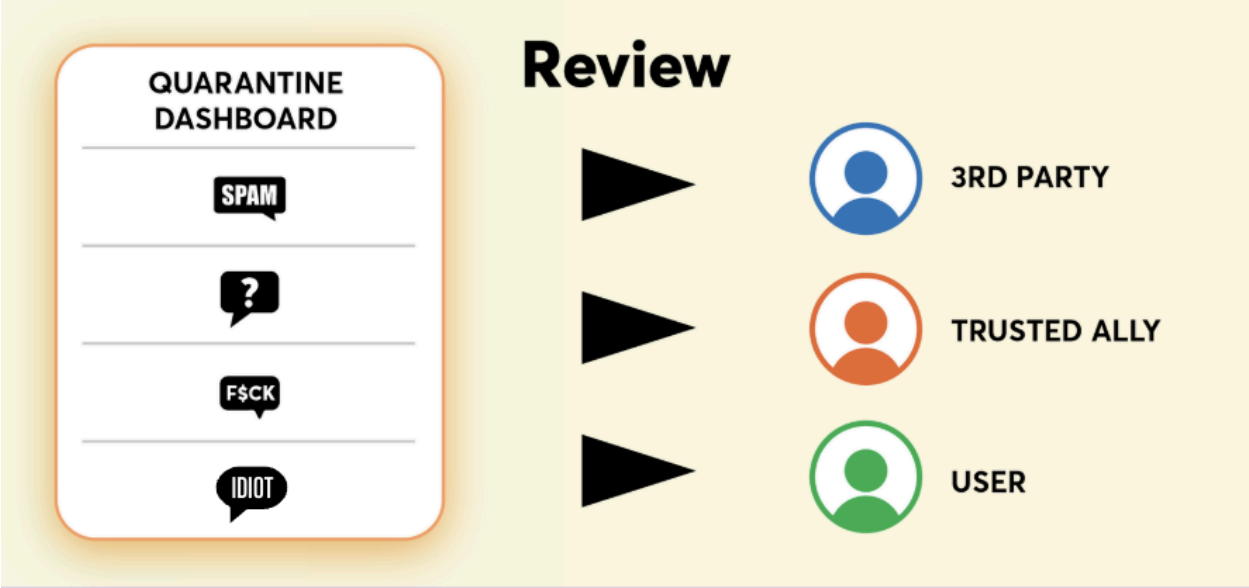


Graphic demonstrating how an online abuse filter could detect and quarantine potentially abusive content in a centralized dashboard. Graphic by Consumer Reports.

Detection

In the detection phase, a filter automatically reviews incoming content, identifies potential abuse, and determines whether such content should be made visible to the impacted user (via their regular feeds, channels, etc.) or hidden from the impacted user and sent to a user-accessible quarantine area instead. The crux of this step is *how* automated filtering would work, which could rely on anything from user-created word filters and community-driven models to personalized, machine learning solutions. We detail what kinds of automated filtering are currently technically feasible, as well as their advantages and challenges, in Section V.

[31] Viktorya Vilk, Elodie Vialle, and Matt Bailey, "No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users," *PEN America,* March 31, 2021, pen.org/report/no-excuse-for-abuse/.



Graphic demonstrating how a user could review potentially abusive content quarantined in a centralized dashboard—or delegate review to a trusted ally. Graphic by Consumer Reports.

Review

In the review phase, content that has been filtered into the quarantined area is organized into a centralized user-facing interface, such as a dashboard. The user can then choose to review all quarantined content in the dashboard and decide what to do with it. In this dashboard, users could also have the option to blur quarantined content by default so that they are not immediately confronted with abuse, but instead opt into seeing it. To further reduce the psychological toll, users could be given the option to "friendsource" the quarantined content by delegating access to their dashboard to a friend or other trusted third party, such as an employer or civil society organization, who could help review and take action.[32]

[32] Kaitlin Mahar, Amy X. Zhang, and David Karger, "Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation," *2018 CHI Conference on Human Factors in Computing Systems*, no. 586 (April 2018), 1-13, homes.cs.washington.edu/~axz/squadbox.html.

Graphic demonstrating what action a user, or their trusted ally, could take on potentially abusive content quarantined in a centralized dashboard. Graphic by Consumer Reports.

Action

In the action phase, the user or their trusted allies could address the abusive content, including marking the content as not abusive to remove it from quarantine and ensure it appears on their feed, timeline, DMs, etc.; reporting the content if it violates platform policies; blocking or muting the accounts behind the abusive content; or otherwise leveraging in-platform features to address it. Even if the user does not take explicit action on the abusive content quarantined in their dashboard, that content has automatically been detected and documented, which is a critical step in enabling targets of abuse to protect themselves and exert agency over their experiences.

To be clear, with this proposed mechanism, potentially abusive content that has been automatically detected and quarantined does *not* automatically disappear from view for all users of the platform; it is hidden only from the targeted user's view via their main feeds, DMs, channels, etc. The targeted user can access all quarantined content at any time, via their dashboard, and remove it from quarantine or otherwise address it as needed.

Section II: Does the Technology to Automatically Detect Online Abuse Currently Exist?

In order to detect spam, abuse, and other forms of harmful content, technology companies rely on a combination of human and automated review and moderation. For automation, platforms rely on AI technologies. AI, in the words of John McCarthy, who coined the term, is "the science and engineering of making intelligent machines, especially intelligent computer programs"[33]—making computer systems that are able to perform tasks commonly associated with human intelligence. Most automated detection systems today that analyze text-based content (whether spam or abuse) rely on some form of Natural Language Processing (NLP)—a specific branch of AI that primarily focuses on giving computers the ability to process and take action on human language.

AI is not new, and neither is NLP. Theoretical interest in AI and NLP emerged in the 1940s and 1950s, and the technologies behind AI and NLP, like deep learning, have existed for a decade.[34] Social media companies are already deploying a range of these technologies to automatically detect harmful content, including online abuse, on their platforms—and have done so for years.[35] Most automated detection technologies are used to facilitate content moderation behind the scenes. The automated detection technology is controlled by employees of platforms or third-party companies rather than being made available to individual users.

Technology companies have gradually integrated features that leverage automated detection to give users more control over their individual experiences. Most social media companies now allow users to manually hide or mute certain kinds of content that they do not want to see in their own feeds, DMs, comments, etc.—though each platform's features are distinct in their functionality and terminology. On most platforms, users can manually set keywords so that responses to their content containing those keywords are automatically concealed under a "hidden replies" cover, which other users then need to click through to reveal the response. This is distinct from allowing users to delete responses to their own content altogether, which makes the content invisible for *all* users.[36] A few platforms have also given users access to a handful of preset filters that reduce the visibility of sensitive or otherwise harmful content. Below we map out the relevant features we could find across multiple major social media platforms:

> **X (formerly Twitter)** allows users to manually mute entire accounts, individual posts, and replies to their posts. Users can also manually mute content by keyword, emoji, or hashtag. They cannot mute DMs, but they can hide DM notifications.[37] In terms of

[33] John McCarthy, "What Is Artificial Intelligence?" *Stanford University Department of Computer Science,* November 12, 2007, formal.stanford.edu/jmc/whatisai.pdf.
[34] Gil Press, "A Very Short History of Artificial Intelligence (AI)," *Forbes,* December 30, 2016, forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/; "Natural Language Processing," *Stanford University,* 2004, cs.stanford.edu/people/eroberts/courses/soco/projects/2004-05/nlp/overview_history.html; Keith D. Foote, "A Brief History of Deep Learning," *DATAVERSITY,* February 4, 2022, dataversity.net/brief-history-deep-learning/.
[35] Rem Darbinyan, "The Growing Role of AI in Content Moderation," *Forbes*, June 14, 2022, forbes.com/councils/forbestechcouncil/2022/06/14/the-growing-role-of-ai-in-content-moderation/; email to PEN America from Reddit spokesperson, March 2025; email to PEN America from Meta spokesperson, January 2023.
[36] Kaya Yurieff, "Twitter Now Lets Users Hide Replies to Their Tweets," *CNN,* November 21, 2019, cnn.com/2019/11/21/tech/twitter-hide-replies/index.html.
[37] "How to mute accounts on X," *X,* accessed March 10, 2025, help.x.com/en/using-x/x-mute; "How to Use Advanced Muting Options," *X,* accessed March 10, 2025, help.x.com/en/using-x/advanced-x-mute-options.

automated filters set by the platform (rather than manual filters set by users), Twitter (now X) introduced a "quality filter" in 2016, which automatically removes "lower-quality content" (e.g., duplicate posts) from users' notifications.[38]

Facebook offers no exact equivalent to X's (formerly Twitter's) manual muting features, but users can "snooze" accounts or groups for 30 days, mute other users' stories, and permanently unfollow posts without unfriending accounts.[39] Users can also select keywords to be blocked from appearing in comments on their profiles.[40] From the standpoint of automated filters set by the platform (rather than manual filters set by users), Facebook allows users to "filter for profanity," but again only on pages and profiles set to Professional Mode rather than for all users.[41]

Instagram, unlike Facebook and X, automatically detects and then hides comments containing offensive words, phrases, and emojis by default, and users have the option to click to reveal the hidden content. Users can manually add additional words or phrases using the "hidden words" feature.[42] Instagram also enables users to manually mute posts or stories and to mute accounts.[43] In terms of platform-set automated filters (rather than user-set manual ones), users can turn on "advanced comment filtering," which will detect and filter additional comments that meet Instagram's criteria for potentially objectionable content.[44] Instagram also introduced a feature called "limits interactions," initially only for influencers, in 2021. Limit interactions allows influencers to automatically limit interactions from recent followers, everyone except close friends, or all users to prevent unwanted accounts from responding to their stories, tagging or mentioning them, commenting on their posts, and remixing their content; in 2024, they made this feature available to all users.[45]

TikTok does not automatically filter or hide abusive content by default. Instead, users can opt into several automated filtering modes: restricted mode, which hides "mature and complex" content, and comment care mode, which hides comments from strangers (people who neither follow a user or are followed by them) and comments that are similar to others that a user has previously disliked, reported, or deleted.[46] Users can manually add keyword filters, which automatically detect and hide comments with the

[38] "About the Notifications Timeline," *X,* accessed August 7, 2024, help.x.com/en/managing-your-account/understanding-the-notifications-timeline; Michael Burgess, "Twitter Introduces 'Quality Filter' to Tackle Harassment, " *WIRED,* August 19, 2016, wired.com/story/twitter-quality-filter-turn-on.
[39] "How Do I Unfollow a Person, Page, or Group on Facebook," *Facebook,* accessed July 18, 2024, facebook.com/help/190078864497547; "Mute or Unmute a Story on Facebook," *Facebook,* accessed July 18, 2024, facebook.com/help/408677896295618.
[40] Email to PEN America from Meta spokesperson, January 2023.
[41] "Manage Comments with Moderation Assist for Pages and Professional Mode," *Facebook,* accessed August 12, 2024, facebook.com/help/1011133123133742; "Turn Comment Ranking On or Off for YourFacebook Page," *Facebook,* accessed July 18, 2024, facebook.com/help/1494019237530934.
[42] "Hide Comments or Message Requests You Don't Want to See on Instagram," *Instagram,* accessed June 24, 2024, help.instagram.com/700284123459336; Email to PEN America from Meta spokesperson, January 2023.
[43] "Mute or Unmute Someone on Instagram," *Instagram,* accessed July 18, 2024, help.instagram.com/469042960409432.
[44] "Hide Comment or Message Requests You Don't Want to See on Instagram," *Instagram,* accessed August 2, 2024, help.instagram.com/700284123459336.
[45] "Temporarily Limit People From Interacting with You on Instagram," *Instagram,* accessed August 7, 2024, help.instagram.com/4106887762741654/?helpref=uf_share.
[46] "Countering Hate Speech and Behavior," *TikTok,* accessed June 25, 2024, tiktok.com/safety/en/countering-hate; "Comment Care Mode," *TikTok,* accessed August 19, 2024, support.tiktok.com/en/safety-hc/account-and-user-safety/comment-care-mode.

keyword(s). Users can also opt in to automatically hiding all comments and then manually approving each comment to make it visible.[47]

The in-platform features outlined above go some way toward enabling users to approach abusive content more like spam, in that users can leverage automated detection technologies to reduce their own exposure. However, there are some important differences in what we are proposing. Currently, to automatically filter content for their own accounts, users often have to choose between hiding/muting content manually or switching on preset filters. In-platform features that allow users to manually set their own keywords, emojis, hashtags, etc., offer more transparency and control but are also more labor-intensive. Most preset filters, on the other hand, are binary and a black box; users can turn the feature either on or off, but they cannot fine-tune it or easily see everything that has been filtered out.[48] All of these features are piecemeal, idiosyncratic, and spread out across multiple different sections of platform apps and websites.

In recent years, several social media companies have been experimenting with significantly more robust, flexible, and transparent mechanisms that allow users to automatically detect and filter out abusive content, which we outline below:

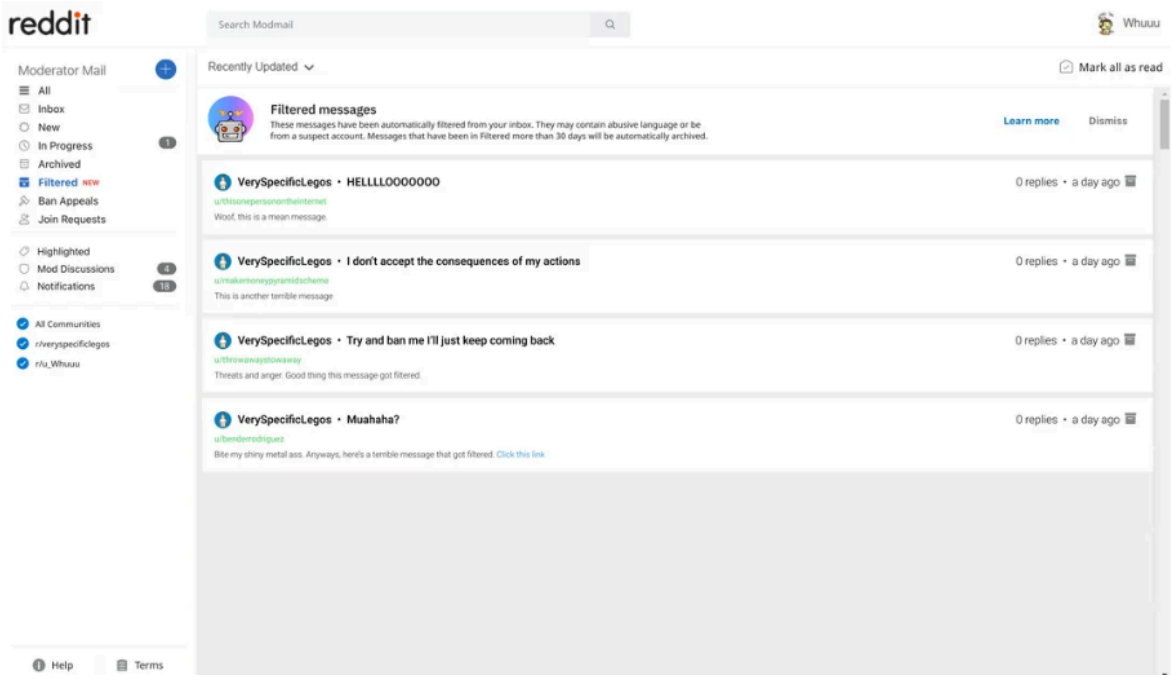
Facebook's Moderation Assist: In 2021, Meta piloted the new Moderation Assist feature on specific parts of Facebook.[49] Page admins or those with profiles in Professional Mode can manually fine-tune a range of criteria to detect and filter out certain content (e.g., if the comment contains a given keyword, or if the commenter doesn't have any friends or followers).[50] Those who use Moderation Assist have access to a Professional Dashboard, where they can review hidden comments and decide how to address them, as well as access their moderation activity log. At present, the Moderation Assist feature is not available to every user on the platform.[51]

YouTube's Channel Comments and Mentions: Since at least 2020, YouTube has allowed users who post videos to review and moderate every comment made in reply to their videos or on their channel. Users choose their own criteria to automatically detect comments and hold them for review before publication. Users can choose to not hold any comments for review, hold "potentially inappropriate comments," hold a broader range of "potentially inappropriate comments," or hold all comments.[52] All incoming comments are then automatically analyzed according to users' settings and divided into two folders. In one folder are comments that the platform has deemed acceptable and automatically published. In the second folder are comments that the platform has held for review, which the user then manually approves or rejects for publication. This system

[47] "Countering Hate Speech and Behavior," *TikTok,* accessed June 25, 2024, tiktok.com/safety/en/countering-hate.
[48] Leigh Honeywell, interview by Yael Grauer, Consumer Reports, December 9, 2021.
[49] Leah Loeb, "Facebook Introduces Comment Moderation and Live Chat for Creators," *Hootsuite,* December 10, 2021, blog.hootsuite.com/social-media-updates/facebook/facebook-comment-moderation-live-chat-for-creators/.
[50] "Manage Comments with Moderation Assist for Pages and Professional Mode," *Facebook,* accessed June 24, 2024, facebook.com/help/1011133123133742.
[51] "Manage Comments with Moderation Assist for Pages and Professional Mode," *Facebook,* accessed June 24, 2024, facebook.com/help/1011133123133742; email to PEN America from Meta spokesperson, January 2023.
[52] "Learn About Comment Settings," *YouTube Help,* accessed July 23, 2024, support.google.com/youtube/answer/9483359?hl=en#zippy=%2Con; Jordan, "Potentially inappropriate comments now automatically held for creators to review," *YouTube Help,* accessed July 23, 2024, support.google.com/youtube/thread/8830320/potentially-inappropriate-comments-now-automatically-held-for-creators-to-review?hl=en.

automatically detects potentially harassing or otherwise unwanted comments, and the user can adjust, to some degree, the strictness of the filter.[53]

Reddit's Harassment Filter and Modmail Folder: Reddit offers multiple harassment detection tools—several specifically to subreddit moderators, rather than to all users.[54] The Harassment Filter automatically detects content that is likely to be considered harassing under the platform's policies. Moderators can then review the flagged content and either approve it to be posted to the community thread or remove it. Moderators are able to refine the Harassment Filter by toggling between filtering fewer comments with more accuracy or filtering more comments with less accuracy. For additional fine-tuning, moderators can also manually add up to 15 keywords that the algorithm will not filter out.[55] In addition, Reddit also offers moderators the Modmail Folder. This feature automatically redirects potentially harassing modmail from a moderator's general inbox to a filtered folder. The Modmail Folder was designed to give moderators greater control over when they view and interact with potentially abusive content. Additionally, they can manually report harassing modmail messages missed by the filter and move into the regular inbox any non-harassing modmail messages that have been erroneously filtered out.[56]



Screenshot of Reddit's Modmail Harassment Filter. Photo by enthusiastic-potato on Reddit.

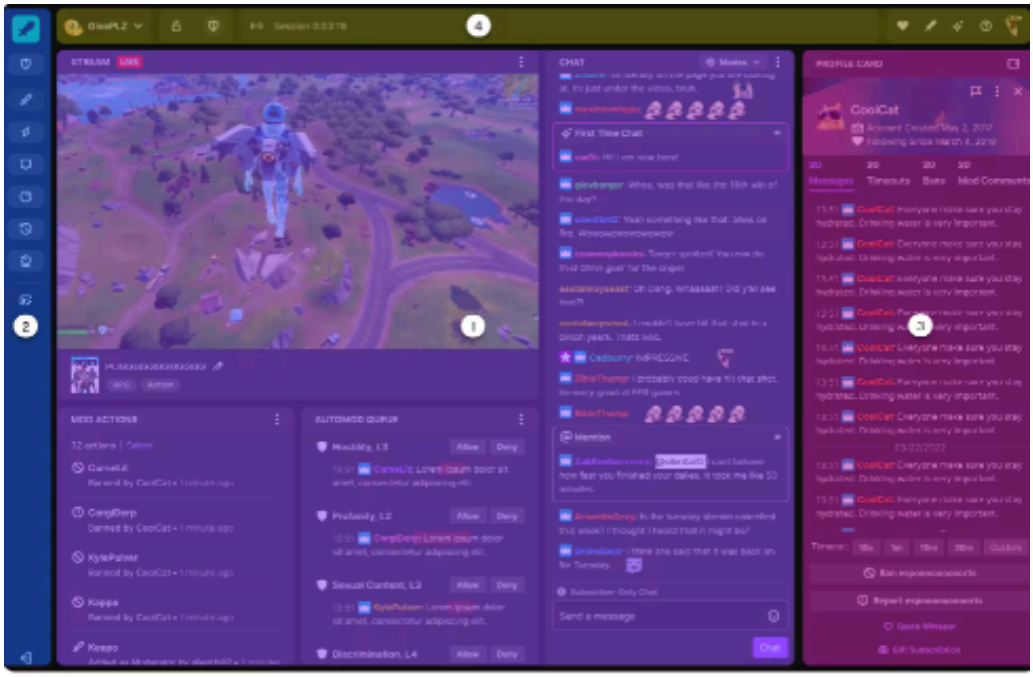
[53] "Review and Reply to Comments," *YouTube Studio App Help Centre,* accessed June 18, 2024, support.google.com/youtubecreatorstudio/answer/9482367?hl=en-IN&co=GENIE.Platform%3DDesktop#:~:text=Check%20comments%20on%20your%20videos%20and%20channel&text=From%20the%20left%20menu%2C%20select,by%20YouTube%20as%20likely%20spam.

[54] Chase DiBenedetto, "Reddit Introduces an AI-powered Tool That Will Detect Online Harassment," *Mashable*, March 7, 2024, mashable.com/article/reddit-ai-harassment-filter; email to PEN America from Reddit spokesperson, March 2025.

[55] "Harassment Filter," *Reddit,* accessed June 25, 2024, support.reddithelp.com/hc/en-us/articles/23856209638932-Harassment-filter.

[56] "Modmail Folders," *Reddit,* accessed June 25, 2024, support.reddithelp.com/hc/en-us/articles/15484158762260-Modmail-folders#h_01G8YBFB9VYVREXYWH1SCDP141; email to PEN America from Reddit spokesperson, March 2025.

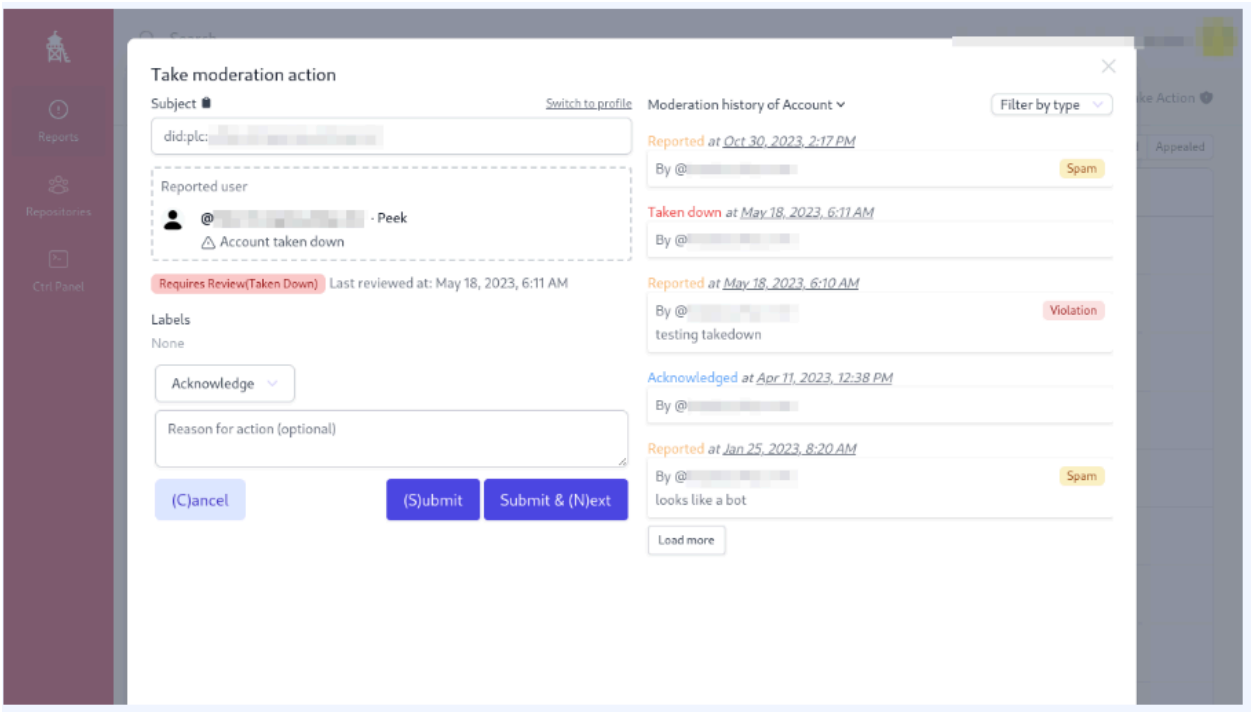
Twitch: Since 2016, Twitch has offered a feature called AutoMod, which uses AI and NLP to flag and quarantine potentially offensive or abusive content shared on livestream chats.[57] Streamers can customize the sensitivity of AutoMod across four levels of severity: Level One, for example, only filters out discrimination, whereas Level Four addresses discrimination, sexual content, profanity, and most forms of hostility.[58] Streamers can then review the quarantined content, and decide to allow or prevent the message to be posted to the public chat. The platform also offers Friendsourcing, where streamers call on friends or fans to help moderate their chats through AutoMod.[59] Streamers can deploy AutoMod in conjunction with other moderation features through Shield Mode, which allows streamers to build a custom set of moderation tools under one centralized shield that can easily be toggled on and off as needed.[60]



Screenshot of Twitch's Mod View features. Photo by Twitch.

[57] Andrew Webster, "Twitch introduces a new automated moderation tool to make chat friendlier," *The Verge,* December 12, 2016, theverge.com/2016/12/12/13918712/twitch-automod-machine-learning-moderation-tool.
[58] "How to Use AutoMod," *Twitch,* accessed February 24, 2025, help.twitch.tv/s/article/how-to-use-automod?language=en_US.
[59] Nicole Carpenter, "Moderators are the unpaid backbone of Twitch," *Polygon,* October 20, 2023, polygon.com/23922227/twitch-moderators-unpaid-labor-twitchcon-2023.
[60] "Protect Your Channel With Shield Mode," *Twitch,* November 30, 2022, safety.twitch.tv/s/article/Protect-your-channel-with-Shield-Mode?language=en_US; email to PEN America from Twitch spokesperson, March 2025.

Bluesky, which launched as an independent entity in 2021, was built from its inception to enable a "stackable" approach to moderation. This means that the platform not only makes use of its own internal moderation teams but also supports the integration of third-party moderation for users. In other words, users who want additional protection from potentially abusive content can set up additional moderation services on top of what Bluesky already offers. Bluesky open-sourced Ozone, its internal moderation tool, in March 2024. This allows individuals and/or teams to set up their own specific preferences to label, triage, and escalate potentially abusive content. Furthermore, Ozone can operate not just on Bluesky but also on any platform that uses the AT Protocol, a communication standard that enables the creation of decentralized social networking platforms; in other words, a user could set up Ozone on Bluesky and then use it across other decentralized platforms.[61]



Screenshot of Bluesky's Ozone tool. Photo by Bluesky.

These newer mechanisms demonstrate that platforms can, in fact, automatically detect and quarantine abusive content. By empowering users to fine-tune how strictly they want to filter content, Bluesky, Twitch and Reddit's features provide particularly helpful models. According to Reddit, takeup rate for and satisfaction with both the Harassment Filter and Modmail Filter are very high among community moderators.[62] Twitch has reported similarly high adoption and

[61] The Bluesky Team, "Bluesky's Stackable Approach to Moderation," *Bluesky,* March 12, 2024, bsky.social/about/blog/03-12-2024-stackable-moderation; Richard Ernszt, "What is AT Protocol (Authenticated Transfer Protocol)?" *Comparitech*, November 25, 2024, comparitech.com/blog/vpn-privacy/what-is-at-protocol/.

[62] In a March 2025 email to PEN America, a Reddit spokesperson reported that 72% of the largest Reddit communities have adopted the Harassment Filter, and 95% have adopted the Modmail Filter, and both have received generally positive feedback from moderators. Many moderators highlight that the filters surface inappropriate content that may otherwise go unnoticed, which helps keep communities safer without increasing the time or energy spent on moderation.

approval ratings of AutoMod and Shield Mode.[63] Bluesky also serves as a useful model because it was set up from the get-go to enable a mix of both internal features and third-party moderation to reduce online abuse.[64] That said, the existing features on most major social media platforms do not provide such flexible and accessible solutions. Only a limited subset of YouTube and Reddit users, for example, can leverage the kinds of fine-tuning features that we advocate for here, and neither platform offers a centralized dashboard in which any individual user can review *all* of the different types of content that has been filtered out on that user's account across *all* of the different parts of the platform (including comments, replies, tags, and DMs).

Because no social media company (or email provider, for that matter) has yet integrated a sufficiently robust, comprehensive, user-friendly mechanism to detect and quarantine abusive content that is available to all users, third-party technology companies have stepped up. These tools have different functions and strengths in allowing individual users to limit their exposure to harassment, but none yet offers the full range of protections we envision.

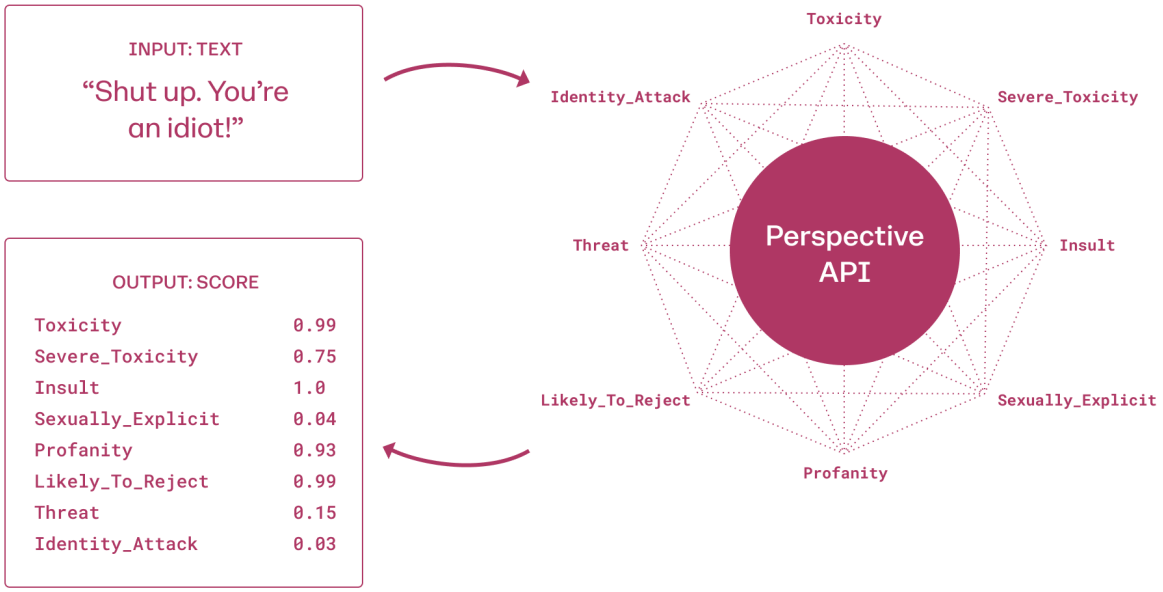
Block Party: In 2021, software engineer and entrepreneur Tracy Chou launched a third-party application that users could plug into Twitter (now X) to reduce exposure to abusive content. Block Party automatically detected abusive content via heuristics rather than machine learning (by, for example, restricting mentions from strangers) and quarantined it into "lockout folders" for later review.[65] Higher tiers of Block Party also gave users the ability to assign helpers to assist with monitoring, muting, or blocking abuse.[66] Unfortunately, this specific Block Party tool is no longer available due to X's ongoing changes to Application Programming Interface (API) access and pricing. Instead, Block Party has pivoted to empowering users to tighten their privacy and safety settings across multiple major social media platforms.[67]

Squadbox: In 2018, MIT's Computer Science and AI Laboratory piloted a thirty-party application that users could plug into their email account to reduce exposure to abusive emails. The platform enabled Friendsourcing: A user could designate a "squad" of supporters to review and manage hateful or harassing email content on their behalf.[68] All emails, aside from those sent by verified contacts, were automatically forwarded to squad supporters who could either quarantine messages or forward them to the targeted individual.[69] The tool is currently in a redevelopment and testing phase.

Perspective API: Launched by Google's Jigsaw and Counter Abuse Technology teams in 2017, Perspective API is a free and open-source Application Programming Interface

[63] In an email to PEN America, a Twitch spokesperson reported that 80% of channels use AutoMod and Shield Mode, up from 20% in 2022.
[64] The Bluesky Team, "The AT Protocol," *Bluesky,* October 18, 2022, bsky.social/about/blog/10-18-2022-the-at-protocol.
[65] Tracy Chou, "Meet the App Developer Creating a Simple Tool That Could Slay All Online Trolls," *BBC Science Focus Magazine,* June 19, 2021, sciencefocus.com/future-technology/bullying-how-to-slay-the-social-media-trolls.
[66] Block Party, blockpartyapp.com/; "Product FAQ," *Block Party,* accessed August 12, 2024, blockpartyapp.com/faq/#what-does-a-helper-do.
[67] Sara Keenan, "Privacy Party Is Block Party's Bold Response To Twitter's API Changes: The New Haven For Online Privacy," *People of Color in Tech,* March 12, 2024, peopleofcolorintech.com/articles/privacy-party-is-block-partys-bold-response-to-twitters-api-changes-the-new-haven-for-online-privacy/.
[68] "Squadbox Team," *Squadbox*, accessed September 3, 2024, squadbox-dev.csail.mit.edu/#team.
[69] Katilin Mahar, Amy X. Zhang, David Karger, "Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation," *CHI 2018: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, no. 586 (April 21, 2018): 1-13, doi/10.1145/3173574.3174160.

(API), which software developers can use to build a tool. It is designed for communications-focused companies, such as media organizations, rather than for individual users. Perspective API can be integrated into the backends of these companies' websites to provide their human moderators with support to more effectively moderate comments and/or to encourage individual users to reconsider posting "toxic" comments.[70] When it launched in 2017, Perspective API used machine learning to automatically detect toxic content, which Jigsaw defined as a "rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion."[71] According to a Jigsaw spokesperson, the newest iterations of Perspective API leverage AI to power the integration of customizable attributes that help community managers configure their own detection rules based on their own norms and guidelines.[72]



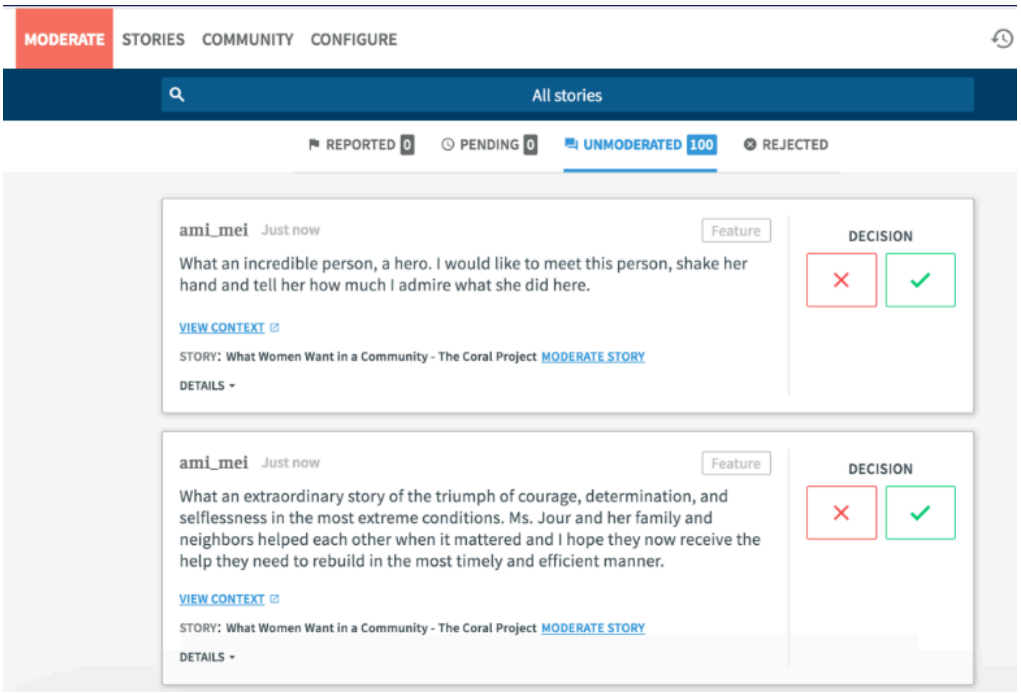
Screenshot demonstrating how Perspective API identifies and evaluates content for toxicity. Photo by Perspective API Developers.

TRFilter: In 2022, the Thomson Reuters Foundation launched TRFilter based on an open source tool built by Google's Jigsaw team called Harassment Manager, which uses Perspective API to help users to automatically detect and document the harassment they received online, built initially for Twitter (now X).[73] This tool enabled journalists to automatically detect and document abusive content on Twitter (now X), as well as to automatically mute, block, and save comments en masse—with light touch manual

[70] "About the API FAQs," *Perspective,* accessed September 3, 2024, support.perspectiveapi.com/s/about-the-api-faqs?language=en_US.
[71] "How It Works," *Perspective,* accessed September 3, 2024, perspectiveapi.com/how-it-works/.
[72] Email to PEN America from Google/Jigsaw spokesperson, March 2025.
[73] "Thomson Reuters Foundation Launches New Tool to Protect Journalists Against Online Violence," *Trust.org,* June 30, 2022, accessed September 4, 2024, trust.org/i/?id=097399b5-492a-4587-89d2-d94d0a0eb259.

review—and to hide abusive content to avoid exposure.[74] Similar to Block Party, TRFilter can no longer function as designed due to X's changes to API access and pricing.[75]

Coral: Coral was initially launched in 2014 as The Coral Project by the Mozilla Foundation, in collaboration with the Knight Foundation, The New York Times, and The Washington Post, and it is now part of Vox Media. Coral is a commenting platform designed for media organizations (rather than individual users) to publish and moderate comments, reviews, and Q&As on their websites and apps.[76] Coral uses many different types of computer intelligence to help with moderation, and offers easy integrations with third-party AI software; one example of this is Coral's optional integration with Perspective API (described above) to calculate a "Toxicity Threshold" that organizations can use to identify and remove abusive comments.[77] While Coral does leverage AI to facilitate automated detection, the tool was always built with the intention of keeping humans in the loop: "We do AI-assisted human moderation," says Andrew Losowsky, who oversees the Coral platform. "Humans are ultimately the main moderators."[78]



Screenshot of Coral's back-end moderation dashboard. Photo by the Coral Project.

[74] Marina Adami and Eduardo Suárez, "Many Journalists Come Under Attack: TRF Launches Free Tool to Monitor Online Abuse," *Reuters Institute for the Study of Journalism*, July 19, 2022, https://reutersinstitute.politics.ox.ac.uk/news/many-journalists-come-under-attack-trf-launches-free-tool-monitor-online-abuse.
[75] David Cohen, "Thomson Reuters Foundation, Google's Jigsaw, Twitter Take Steps to Protect Journalists," *Adweek*, June 30, 2022, adweek.com/media/thomson-reuters-foundation-googles-jigsaw-twitter-take-steps-to-protect-journalists/; Ester Sun, "Digital Tool Helps Shield Journalists from Online Violence," *Voice of America,* July 15, 2022, voanews.com/a/digital-tool-helps-shield-journalists-from-online-violence/6660881.html.
[76] "The Coral Project Is Moving to Vox Media," *Mozilla Blog,* January 22, 2019, blog.mozilla.org/en/mozilla/the-coral-project-is-moving-to-vox-media/#:~:text=Since%202015%2C%20the%20Mozilla%20Foundation,%2Dcentered%2C%20open%20source%20software.
[77] "Toxic Comments," *The Coral Project*, accessed September 4, 2024, legacy.docs.coralproject.net/talk/toxic-comments/.
[78] Andrew Losowsky, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 5, 2022.

While each of these third-party tools approximate the idea of treating online abuse like spam, none do exactly what we propose. Two of them, Perspective API and Coral, are built for use by organizations rather than individuals. Of the three built for individual use—Block Party, Squadbox, and TRFilter—none are currently operational as originally intended.

What we propose is that platforms integrate a mechanism that combines aspects of all of the in-platform features and third-party tools described above in order to provide users with greater control over the content they see. Platforms can build this mechanism themselves or they can ensure free and open access to their API so that third-party companies can build such a mechanism; what is important, however, is that the mechanism be secure and directly and smoothly integrated into the platform's primary user experience. We offer detailed recommendations for how to build such a mechanism in Section VII. New technologies like large language models (LLMs) and generative AI (GenAI) are a significant advancement in the field of NLP and could be further harnessed to improve automated abuse detection, which we explore in the next section.

Section III: What Role Can LLMs and GenAI Play?

LLMs and GenAI—two relatively new technologies that exploded in popularity since the release of ChatGPT in November 2022—represent a significant advancement in the capabilities of AI to analyze existing content and to generate new content.[79] It is therefore important to consider how these technologies may impact both the prevalence of online abuse and its detection.

LLMs are the most current version of NLP systems that can effectively analyze and generate human language text. At a high level, the way LLMs work is by scanning enormous amounts of available text, usually from across the web—including news articles and social media—and using that data to *train* the algorithms that power the model. The model can then, when fed new information, analyze text, make predictions, or even generate new text. LLMs excel at pattern recognition and are rapidly becoming more sophisticated, powering chatbots like ChatGPT.[80]

Unlike traditional AI systems that are designed to recognize patterns and make predictions, GenAI can generate *new* content in response to prompts—in the form of text, images, audio, videos, computer code, and beyond. Like LLMs, GenAI relies on models trained on enormous amounts of data, mostly from the web.[81] Over the past two years, GenAI has become significantly more convincing, affordable, and accessible, powering systems like DALL-E.

In PEN America's 2023 report, "Speech in the Machine," the authors took the view that AI technologies are not inherently good or bad for free expression, but that "what matters is who uses them, how they are being used, and what stakeholders can do to shape a future in which new technologies support and enhance fundamental rights." This includes the social media sphere. On the one hand, LLMs and GenAI can be harnessed to supercharge disinformation and online abuse campaigns because they can generate more text, audio, and video content that looks and sounds more convincingly human, in more languages, faster. On the other hand,

[79] Bernard Marr, "A Short History Of ChatGPT: How We Got To Where We Are Today," *Forbes*, May 19, 2023, forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/.
[80] "What are Large Language Models (LLMs)?" *IBM,* accessed September 11, 2024, ibm.com/topics/large-language-models.
[81] Nick Routley, "What is generative AI? An AI explains," *World Economic Forum,* February 6, 2023, weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/.

these technologies can also be used to facilitate the detection of abusive or otherwise harmful content in order to more effectively and efficiently address it.[82]

Researchers have started exploring how LLMs and GenAI could help proactively detect harmful content. Recent research from Stanford University, UC San Diego, and the University of Buffalo demonstrated that LLMs can achieve significantly better accuracy in automatically detecting abusive content online than state-of-the-art non-LLM models, especially in the context of Standard American English.[83] The model performed better at automatically detecting content that violates platform policies when it was provided with the full context of a conversation, which represents a significant step forward. This early research suggests that automated systems to detect abuse might soon benefit from LLMs in specific contexts. In fact, Reddit has already incorporated LLMs into its Harassment Filter for moderators (for more, see Section II), and a handful of moderators are already reporting improved performance for specific categories of harm.[84]

Nevertheless, LLMs still have significant room for improvement before they can effectively facilitate automated abuse detection across a broader range of contexts, languages, and types of media because of implicit bias and other challenges (discussed in detail in Section V). They are also still prohibitively expensive to train, let alone deploy, maintain, and retrain to ensure accuracy over time, which poses a significant barrier to their widespread adoption.[85] So while LLMs and GenAI have the potential to supercharge online abuse, as well as to improve its automated detection, the degree to which they will be helpful or harmful remains to be seen.

Section IV: Why Don't Platforms Do More to Reduce Exposure to Online Abuse?

The technology to treat online abuse more like spam exists. While detecting and filtering abuse is considerably more challenging than detecting and filtering spam—as discussed in detail in the next section—it is doable. In fact, platforms have been using the baseline technology that enables the automated detection of potentially harmful content for years for their own internal content moderation systems. Platforms like YouTube, Reddit, Twitch, and Bluesky, as outlined in Section II, are experimenting with giving users access to limited aspects of such technology, but no platform has yet built a robust, comprehensive mechanism akin to the one we propose.

Most of the sources we interviewed agreed that taking inspiration from spam to reduce exposure to online abuse was well worth trying. Given that this idea has been around for at least seven years, why haven't some of the wealthiest technology companies—which are routinely criticized for inadequately addressing abuse on their platforms—implemented it?

---

[82] Summer Lopez, "Speech in the Machine: Generative AI's Implications for Free Expression," *PEN America,* July 21, 2023, pen.org/report/speech-in-the-machine/.
[83] Deepak Kumar, Yousef AbuHashem, Zakir Durumeric, "Watch Your Language: Investigating Content Moderation with Large Language Models," *ArXiv,* September 25, 2023, arxiv.org/pdf/2309.14517; Kenyan Guo et al., "An Investigation of Large Language Models for Real-World Hate Speech Detection," *ArXiv,* January 7, 2024, arxiv.org/pdf/2401.03346.
[84] "Harassment Filter," *Reddit,* support.reddithelp.com/hc/en-us/articles/23856209638932-Harassment-Filter; email to PEN America from Reddit spokesperson, March 2025.
[85] Craig Smith, "What Large Models Cost You – There Is No Free AI Lunch," *Forbes*, January 1, 2024, forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch.

Building robust anti-harassment mechanisms—particularly proactive detection and quarantine measures—requires significant investment of time, energy, money, and staff. According to nearly all of our sources, many of whom have worked in the technology industry, investing in tools to alleviate online abuse is simply not a high priority for most major technology companies. "You need people [i.e., staff] to solve this problem and you need to really invest in it," says Caroline Sinders, founder of human rights design agency Convocation Design + Research. "And companies don't seem to want to do that."[86]

Indeed, most major technology companies added abuse mitigation mechanisms only retroactively, after years of intense activism and public pressure. As mentioned in the introduction, Twitter (now X) did not integrate a reporting button until 2013, seven years after the app launched in 2006, and only after public outcry.[87] In 2017, more than 13 years after it was created, Facebook finally allowed users to quarantine direct messages sent from abusive accounts without having to block the abusive user;[88] the platform only allowed users to report abuse on someone else's behalf in 2018.[89] When Instagram launched in 2010, users could report abuse only through a separate form and were advised to manually delete abusive comments;[90] it did not introduce a mute button until 2018.[91] Twitch launched Shield Mode, a customizable safety feature, only after prominent livestreamers staged a virtual walkout in response to prolific "hate raids"—extreme, coordinated abuse campaigns—that had occurred a year prior.[92]

The negligence of technology companies runs deeper than indifference. Many researchers have made the case that the *incentives* to meaningfully reduce the amount of hateful and harassing content on social media platforms are misaligned. There is broad consensus that user engagement is a critical component to platform success because it drives ad revenue. Researchers have highlighted how technical architecture, which is fundamentally focused on boosting engagement, can increase abusive behaviors online.[93] The result of such incentives, Kent Bausman, PhD, a sociology professor at Maryville University in St. Louis, said in a 2020 Forbes article, is that social media "has made trolling behavior more pervasive and virulent."[94]

---

[86] Caroline Sinders, interview by Yael Grauer, Consumer Reports, December 14, 2021.
[87] Alexander Abad-Santos, "Twitter's 'Report Abuse' Button Is a Good, But Small, First Step," *The Atlantic*, July 31, 2013, theatlantic.com/technology/archive/2013/07/why-twitters-report-abuse-button-good-tiny-first-step/312689/; Abby Ohlheiser, "The Woman Who Got Jane Austen on British Money Wants To Change How Twitter Handles Abuse," *Yahoo News,* July 28, 2013, news.yahoo.com/woman-got-jane-austen-british-money-wants-change-024751320.html.
[88] Antigone Davis, "New Tools to Prevent Harassment," *About Facebook,* December 19, 2017, about.fb.com/news/2017/12/new-tools-to-prevent-harassment/.
[89] Antigone Davis, "Protecting People from Bullying and Harassment," *About Facebook,* October 2, 2018, about.fb.com/news/2018/10/protecting-people-from-bullying/.
[90] "User Disputes," *Wayback Machine,* October 18, 2011, accessed February 16, 2021, web.archive.org/web/20111018040638/help.instagram.com/customer/portal/articles/119253-user-disputes.
[91] Megan McCluskey, "Here's How You Can Mute Someone on Instagram Without Unfollowing Them," *TIME*, May 22 2018, time.com/5287169/how-to-mute-on-instagram/.
[92] Ash Parrish, "After weeks of hate raids, Twitch streamers are taking a day off in protest," *The Verge,* August 31, 2021, theverge.com/2021/8/31/22650578/twitch-streamers-walkout-protest-hate-raids; Jon Fingas, "Twitch's new 'Shield Mode' is a one-button anti-harassment tool for streamers," *Engadget,* November 30, 2022, engadget.com/twitch-shield-mode-anti-harassment-180003661.html.
[93] Luke Munn, "Angry by Design: Toxic Communication and Technical Architecture," *Humanities and Social Sciences Communications* 7, no. 53 (July 2020): doi.org/10.1057/s41599-020-00550-7.
[94] Peter Suciu, "Trolls Continue to be a Problem on Social Media," *Forbes*, June 4, 2020, forbes.com/sites/petersuciu/2020/06/04/trolls-continue-to-be-a-problem-on-social-media/.

Several of our sources concurred, arguing that because publicly held companies are driven by increasing profit for shareholders, they are unlikely to engage in steps that reduce online abuse if creating a safer, less toxic space for users would significantly impact their bottom line. "These companies are for profit companies," says Susan McGregor, a research scholar at Columbia University's Data Science Institute. "Most of them are publicly held. They have a legal obligation to maximize their profitability in the absence of regulation to the contrary … to protect their shareholders' interest. … Speech that is abusive and harassing towards an individual or a group on social media is abuse and harassment, as far as the targets are concerned. To platforms that host it, it looks like engagement, and engagement equals advertisers."[95]

We encountered a different perspective from trust and safety professionals who have worked in the tech industry to develop and enforce policies that define acceptable behavior and content online.[96] Many trust and safety experts argue that effective content moderation and anti-abuse mechanisms *should* be aligned with platforms' growth incentives. The Trust & Safety Professional Association, for example, states, "Content moderation is also a required business investment to ensure the desired traffic and growth to the platform…. as it directly influences user trust and brand reputation, which in turn influences growth and revenue."[97]

Some research also suggests that a safer, more inclusive online environment is more sustainably profitable than an economy of abuse. A 2018 report found that women were 26% less likely to use the internet than their male counterparts, with many women citing safety concerns, including online abuse, as a reason they limit their internet usage.[98] A 2023 study by the National Democratic Institute found that efforts to create a more safe and inclusive environment, particularly for marginalized groups like women, can lead to better reputation and greater brand loyalty for a company. According to the study, some companies, like Bumble, have taken harm reduction more seriously because it is tied directly to brand and overall financial outcomes.[99] "In the context of commercial platforms, revenue and growth pressure is the primary driver … and those pressures lead towards moderation," says Yoel Roth, technology policy fellow at UC Berkeley and formerly head of Trust and Safety at Twitter (now X). "Advertisers are a profoundly influential force on the decisions that ad-supported platforms make. If your goal is safety, advertisers are pretty much on the same page and have … actually pushed companies to do more."[100] In other words, too much abusive and harmful content may drive both users and advertisers away from the platform, therefore undercutting companies' growth and profit goals.

We have seen this effect borne out recently with X. In October 2022, tech entrepreneur Elon Musk bought Twitter and soon renamed the platform X. Under the guise of saving money and protecting free speech, X, under the leadership of Musk, drastically cut the company's Trust and Safety staff, restored alt-right accounts that had previously been banned for violating platform

[95] Susan McGregor, interview by Yael Grauer, Consumer Reports, December 2, 2021.
[96] "About Us," *Trust & Safety Professional Association,* October 14, 2022, tspa.org/about-tspa/.
[97] "What is Content Moderation?" *Trust & Safety Professional Association,* September 19, 2022, tspa.org/curriculum/ts-fundamentals/content-moderation-and-operations/what-is-content-moderation/.
[98] Oliver Rowntree, "The Mobile Gender Gap Report 2018," *GMSA,* Febuary 2018, gsma.com/solutions-and-impact/connectivity-for-good/mobile-for-development/wp-content/uploads/2018/04/GSMA_The_Mobile_Gender_Gap_Report_2018_32pp_WEBv7.pdf.
[99] Theodora Skeadas and Kaleigh Schwalbe, "Technology Companies Must Make Platforms Safer for Women in Politics," *Technology Policy Press,* August 22, 2023, techpolicy.press/technology-companies-must-make-platforms-safer-for-women-in-politics/.
[100] Yoel Roth, interview by Yael Grauer, Consumer Reports, and Deepak Kumar, PEN America and University of California San Diego, December 19, 2023.

policies on hate and harassment, and weakened the platform's popular block feature.[101] Hate and harassment drastically spiked on the platform; studies from the Center for Countering Digital Hate have shown that X failed to remove 86% of tracked hate speech posts, that posts involving anti-Black slurs rose a staggering 202%, and that anti-LGTBQ+ extremists gained followers at quadruple the pre-Musk rate.[102] X has since seen a 59% decrease in advertising revenue because a significant number of advertisers became increasingly reluctant to buy ads on the platform in part out of fear that their brands would appear next to hateful or harassing content that could tarnish their reputation.[103]

Technology companies have historically underinvested in trust and safety teams and are increasingly downsizing them further. X, for example, has cut 43% of its Trust and Safety staff since 2022.[104] As Anika Navaroli, senior fellow at Columbia Journalism School and former senior policy official at Twitter (now X) and Twitch, said, "Even when trust and safety was at its heyday, we still didn't have the investment and the resources for folks to be able to say, we're going to take this engineering team off of building this product that we think is going to bring in all of this revenue, so that you can think about abuse. … it doesn't get to the bottom line."[105] This is underscored by the fact that platforms regularly insist that their trust and safety teams calculate and demonstrate a positive return on investment (ROI) to justify the cost of programs and features to reduce harm, which many trust and safety professionals have noted is a difficult calculation and may not always offset the costs of developing robust moderation and safety systems.[106]

During the tech recession of 2023, major companies like Meta, Amazon, Alphabet, and X (formerly Twitter) all drastically cut the size of their trust and safety teams further.[107] Despite the fact that online abuse, rapidly accelerated by developments in technology like generative AI, is more urgent than ever, major technology companies continue to deprioritize the issue. As Theodora Skeadas, chief of staff at Humane Intelligence and former associate on public policy at Twitter (now X), said: "The layoffs speak for themselves. It shows where the priorities are."[108]

---

[101] Jyoti Mann, "Layoffs, Long Hours and RTO: All the Changes and Controversies in the 12 Months Since Elon Musk Bought X," *Business Insider*, July 27, 2023, web.archive.org/web/20240102033411/https://www.businessinsider.com/elon-musk-twitter; Chris Vallance and Shayan Sardarizadeh, "Tommy Robinson and Katie Hopkins Reinstated on X," *BBC,* November 6, 2023, bbc.com/news/technology-67331288; Robert Hart, "Elon Musk Is Restoring Banned Twitter Accounts–Here's Why the Most Controversial Users Were Removed and Who's Already Back," *Forbes,* August 18, 2023, forbes.com/sites/roberthart/2022/11/25/elon-musk-is-restoring-banned-twitter-accounts-heres-why-the-most-controversial-users-were-suspended-and-whos-already-back/; Emma Roth and Kylie Robison, "X will let people you've blocked see your posts," *The Verge,* September 23, 2024, https://www.theverge.com/2024/9/23/24252438/x-blocked-users-view-public-posts.

[102] "The Musk Bump: Quantifying the rise in hate speech under Elon Musk," *Center for Countering Digital Hate,* December 6, 2022, counterhate.com/blog/the-musk-bump-quantifying-the-rise-in-hate-speech-under-elon-musk/; "X Content Moderation Failure," *Center for Countering Digital Hate,* September 2023, counterhate.com/research/twitter-x-continues-to-host-posts-reported-for-extreme-hate-speech/.

[103] Ryan Mac and Tiffany Hsu, "Twitter's U.S. Ad Sales Plunge 59% as Woes Continue," *The New York Times,* June 5, 2023, nytimes.com/2023/06/05/technology/twitter-ad-sales-musk.html.

[104] Ben Goggin, "Big Tech companies reveal trust and safety cuts in disclosures to Senate Judiciary Committee," *NBC News*, March 29, 2024, nbcnews.com/tech/tech-news/big-tech-companies-reveal-trust-safety-cuts-disclosures-senate-judicia-rcna145435.

[105] Liz Lee and Anika Navaroli, interview by Deepak Kumar, PEN America and University of California San Diego, May 7, 2024.

[106] Alice Hunsberger, "Don't fall into the T&S ROI trap," *Everything in Moderation*, April 8, 2024, everythinginmoderation.co/trust-safety-roi-trap/.

[107] Hayden Field, "Tech Layoffs Ravage the Teams That Fight Online Misinformation and Hate Speech," *CNBC,* May 26, 2023, cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html.

[108] Theodora Skeadas, interview by Deepak Kumar, PEN America and University of California San Diego, January 5, 2024.

Technology companies often invest in robust trust and safety measures only if they are incentivized by public pressure, government regulation, and impact on their bottom line to do so.

Section V: What Are the Challenges With Automatically Detecting Online Abuse?

Spam and online abuse have numerous similarities. For both, the targeted user receives content that is essentially unwanted. And both come with legitimate harms: Spam primarily poses a financial or cybersecurity threat, and online abuse can have serious psychological impacts and can even undermine physical safety.[109] However, there are important differences that make abuse more complicated than spam to automatically detect and quarantine. Below we discuss a range of challenges revolving primarily around context, bias, and the ability of models to keep up with the speed that human language evolves. It is important to note that most of these challenges are inherent to automated detection technology, whether deployed exclusively by platforms behind the scenes to facilitate content moderation or also provided to individual users to empower them to shape their own online experiences.

More Complex Detection Signals

Technology companies, from email providers to social media companies, use multiple different signals to automatically detect both spam and abuse. These signals rely not only on analyzing the *content* of a message but also on *metadata*, which is data *about* the message, such as email headers (which spammers often forge) or IP addresses (a unique string of numbers that identifies each device on the internet).

Volume, for example, is a particularly useful signal for spam because spammers will utilize hundreds of thousands of bots to coordinate large-scale campaigns to target as many people as possible.[110] As spam tends to operate at scale, defenses against spam are also deployed widely; there are multiple companies, for example, that track the specific IP addresses and domain names (web addresses) associated with spam and that provide technology companies with up-to-date block lists. Finally, the ubiquity of spam enables "collaborative" filters, which can automatically flag a message as spam for all users once a critical mass of users have manually marked that message as spam.[111]

Online abuse is more multifaceted than spam, so signals that rely on volume and scale do not work as well.[112] Accounts that spam often typically post *only* spam, whereas accounts that abuse can also post non-abusive content. In a recent paper studying online harassment on

[109] HRC Staff, "New Human Rights Campaign Foundation Report: Online Hate & Real World Violence Are Inextricably Linked," *Human Rights Campaign,* December 13, 2022, hrc.org/press-releases/new-human-rights-campaign-foundation-report-online-hate-real-world-violence-are-inextricably-linked.
[110] Brett Stone-Gross et al. "The underground economy of spam: a botmaster's perspective of coordinating large-scale spam campaigns," *4th USENIX Conference on Large-scale Exploits and Emergent Threats*, (March 2011), 1-4, usenix.org/legacy/event/leet11/tech/full_papers/Stone-Gross.pdf.
[111] Saadat Nazirova, "Survey on spam filtering techniques," *Communications and Network* 3, no. 3 (August 2011): 153-160, dx.doi.org/10.4236/cn.2011.33019.
[112] Malena Dailey, "By the Numbers: What Content Social Media Removes and Why," *Netchoice,* netchoice.org/wp-content/uploads/2021/11/Content-Moderation-By-The-Numbers-v5.pdf.

Reddit, researchers found that only 2.9% of comments posted by accounts that engage in abusive behavior are toxic.[113] "There is a huge difference in the number of messages sent by [a] mailbox that's originating spam messages and the mailbox of your mom," says Laura Edelson, PhD, assistant professor of computer science at Northeastern University in Boston. "These are fairly easy to tell apart. … Harassment has a different distribution. Toxicity is concentrated in a small number of people … [who] send a lot of harassing and toxic messages. Then there's a medium-sized group of people who send one in their lives."[114]

Because the signals used to detect online abuse are often driven by the content itself, rather than by metadata[115]—and because automatically analyzing content can be more difficult than analyzing metadata—detecting abuse is more challenging than detecting spam. As Sinders says: "It's a lot lighter of an analysis to do spam versus trying to do contextual, natural language based processing on harassment because toxicity can look like so many different kinds of things."[116]

Harder to Define

Online abuse is significantly harder to define than spam. "There is a generally agreed definition of what spam is across all platforms, and what its goals are," says Andrew Losowsky, head of community product at Vox Media. "That's not the case for abuse."[117] Spam is defined as unsolicited messages sent via the internet to large groups of people, typically for fraud (e.g., financial scams) or for commercial purposes (e.g., advertising).[118] There is often a mechanism in the spam delivery that is intended to entice and dupe the user into taking an action—usually a hyperlink—which can serve as a useful signal for detecting spam.[119] Users and platforms tend to view spam as binary, categorizing messages as either spam or not spam.[120] Platforms take more consistent action on spam, compared with abuse, in part because there is clearer alignment between platforms and users about how spam is defined and how best to address it. McGregor says: "Spam is expensive and annoying for email providers, and it is annoying and harmful for users."[121]

By contrast, there is currently no widely agreed upon definition of online abuse. In fact, different organizations use different terms (online abuse, online harassment, cyberharassment, online violence, etc.) and define those terms in different ways. PEN America, for example, uses the terms "online abuse" and "online harassment" interchangeably and defines these terms as the "pervasive or severe targeting of an individual or group online through harmful behavior," a

[113] Deepak Kumar et al., "Understanding the Behavior of Toxic Accounts on Reddit," *Association for Computing Machinery,* April 30, 2023, dl.acm.org/doi/abs/10.1145/3543507.3583522.
[114] Laura Edelson, interview by Deepak Kumar, PEN America and University of California San Diego, and Viktorya Vilk, PEN America, May 21, 2024; also mentioned in interview with Caroline Sinders, interview by Yael Grauer, Consumer Reports, December 14, 2021.
[115] Andrew Losowsky, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 5, 2022.
[116] Caroline Sinders, interview by Yael Grauer, Consumer Reports, December 14, 2021.
[117] Andrew Losowsky, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 5, 2022.
[118] "Spam," *Merriam-Webster Dictionary*, merriam-webster.com/dictionary/spam.
[119] Chris Grier et al., "@spam: the underground on 140 characters or less," *17th ACM Conference on Computer and Communications Security*, October 2010, 27-37, dl.acm.org/doi/abs/10.1145/1866307.1866311.
[120] Chris Kanich et al., "Spamalytics: An Empirical Analysis of Spam Marketing Conversion," *Communications of the ACM* 52, no. 9 (September 2009), 99-107, icir.org/christian/publications/2008-ccs-spamalytics.pdf.
[121] Susan McGregor, interview by Yael Grauer, Consumer Reports, December 2, 2021.

deliberately broad definition encompassing a wide range of adversaries and tactics.[122] The nonprofit Electronic Frontier Foundation uses the term "online harassment" and defines it as circumstances under which users find themselves "enduring extreme levels of targeted hostility, often accompanied by the exposure of their private lives."[123] UN Women uses the term "technology-facilitated gender-based violence," which it defines as "any act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms."[124] Consumer Reports' Security Planner describes "online abuse and harassment" as including tactics such as "hateful slurs, unwelcome sexual advances, violent threats, and the unwanted release of personal information."[125] And Google's Jigsaw team, which originally built anti-harassment technology like Perspective API (described in Section II), uses the term "toxicity" (rather than "online abuse"), which it defines as a "rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion."[126]

Context-Dependent

While spam is essentially binary, online abuse is often in the eye of the beholder. Whether or not something is experienced as abusive often depends on the targeted individual's perception, which is rooted in their identity and experiences.[127] Conversations online can be inflected by users' interpersonal relationships.[128] Leigh Honeywell, the CEO at online safety and security company Tall Poppy, stresses this central tension in dealing with online abuse. "There are contexts in which a reply that has a swear word is just friendly banter between actual friends. And there's a context in which that's harassment."[129] Beyond individual perception and interpersonal context, cultural and societal context also play a critical role in determining whether people consider certain types of content abusive. As Edelson explained, "The way people make threats, the way people harass other people, they differ by language and culture."[130]

Intent is critically important but particularly difficult for automated systems to ascertain. While it is certainly possible for a well-intentioned message to be misinterpreted as abuse, it is more likely that some abusers are sophisticated enough to claim good (or neutral) intent when

[122] "Defining 'Online Abuse': A Glossary of Terms," Online Harassment Field Manual, *PEN America,* accessed March 10, 2025, onlineharassmentfieldmanual.pen.org/defining-online-harassment-a-glossary-of-terms/.
[123] Danny O'Brien and Dia Kayyali, "Facing the Challenge of Online Harassment," *Electronic Frontier Foundation,* January 8, 2015, eff.org/deeplinks/2015/01/facing-challenge-online-harassment.
[124] "Repository of UN Women's Work on Technology-Facilitated Gender-based Violence (October 2024)," *UN Women,* October 2024, unwomen.org/sites/default/files/2024-10/repository-of-un-womens-work-on-technology-facilitated-gender-based-violence-en.pdf.
[125] "Get Help Responding to Online Harassment," *Consumer Reports Security Planner,* accessed April 15, 2025, securityplanner.consumerreports.org/tool/get-help-with-online-harassment.
[126] "Attributes & Languages," *Perspective | Developers,* Accessed March 10, 2025, developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US.
[127] Lora Aroyo et al., "Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions," *2019 World Wide Web Conference*, May 2019, 1100-1105, dl.acm.org/doi/abs/10.1145/3308560.3317083.
[128] Email to PEN America from Meta spokesperson, January 2023: "Words and phrases (e.g., 'lol you slut') could mean very different things depending on the relationship, if any, between the sender and receiver, which we often cannot infer…Since bullying and harassment is highly personal by nature, using technology to proactively detect these behaviors can be more challenging than other types of violations."
[129] Leigh Honeywell, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, December 9, 2021.
[130] Laura Edelson, interview by Deepak Kumar, University of California San Diego, and Viktorya Vilk, PEN America, May 21, 2024.

purposely wording their message to be experienced as abusive. The image of an empty egg carton—which is frequently levied by alt-right trolls toward female journalists—is just one example. At its face, the image might seem harmless, but the women who receive these messages understand the implication: The image is an attack on their fertility and life choices.[131]

Implicit Bias

The AI models that power automated detection are known to have a problem with bias because it is often baked into the data used in their design, development, and training.[132] Training an AI model to analyze new content requires pulling together a large body of existing content (data) and then having either human beings or machines annotate that data. While there is a proliferation of models on the market today, many of the most widely used and publicly available models are trained on comments from platforms like Twitter and Reddit ... [which] tend to be white male dominated," says Jacqueline Comer, founder and chief product officer at Areto Labs. "So it's language from only one segment of society that you're training your models on. And then the people labeling those data sets are probably not representative of society either."[133] In a recent paper, researchers at three universities showed that the beliefs and identities of the individuals who annotate training data can inject bias into tools that detect toxic language, highlighting that the data underpinning many hate speech detection models can be biased prior to the development of any technology at all.[134]

Biases built into automated systems disproportionately impact groups that are already underrepresented or marginalized for their identities. For example, researchers have documented the risk of racial bias in detecting hateful speech, showing that hateful speech classification tools often misclassify content written in African American English as hateful when compared with content written in Standard American English.[135] Research conducted in 2017 to 2018 by Google's Jigsaw team, the creators of Perspective API, found that machine learning models erroneously flagged the phrase "I am a gay man" as toxic. This occurred because many uses of the word "gay" in the training data were negative in nature (e.g., used to mock, ridicule, or harass). The model learned a *negative* association between the word "gay" and the toxicity of a message broadly, rather than learning the specific nuance of how the word was ultimately used. In order to solve the issue, the Jigsaw team had to retrain their model on many "counterbalanced" examples of *positive* sentiment content that also included the word "gay."[136]

---

[131] Vicky Spratt, "Why Are Members Of The Alt-Right Sending Me Pictures Of Empty Egg Boxes?" *Refinery 29*, March 6, 2020, refinery29.com/en-gb/2020/03/9512337/anti-feminism-alt-right-fertility.

[132] IBM Data and AI Team, "Shedding Light on AI Bias with Real-World Examples," *IBM*, October 16, 2023, ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples.

[133] Jacqueline Comer, interview by Yael Grauer, Consumer Reports, November 30, 2021.

[134] Maarten Sap et al., "Annotators with Attitudes: How Annotator Beliefs and Identities Bias Toxic Language Detection," *ArXiv,* May 9, 2022, arxiv.org/pdf/2111.07997; similar point made in interview with Kat Lo, by Deepak Kumar, PEN America and University of California San Diego, and Viktorya Vilk, PEN America, December 15, 2021.

[135] Thomas Davidson, Debasmita Bhattacharya, Ingmar Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets," *Association for Computational Linguistic Anthology, Proceedings of the Third Workshop on Abusive Language Online* (August 2019): 25–35, aclanthology.org/W19-3504/.

[136] Luxas Dixon et al., "Measuring and Mitigating Unintended Bias in Text Classification," *Google,* December 27, 2018, storage.googleapis.com/gweb-research2023-media/pubtools/pdf/ab50a4205513d19233233dbdbb4d1035d7c8c6c2.pdf.

Multiple Languages and Multiple Media

As discussed above, most AI systems that automatically detect abusive content rely largely on analyzing the text of the content itself. The problem is that AI systems are not equally good at analyzing text-based content across different languages. In fact, they are significantly better at analyzing content in Standard American English than in other languages. Recent research from the Center for Democracy and Technology identified critical gaps in many LLMs' ability to perform tasks in "low-resource" languages, where training data is limited because the language is not as widely used, studied, or taught online as Standard American English;[137] at present, Somali, Amharic, and Uyghur are examples of low-resource languages.[138] While these LLMs will likely continue to improve, at least one expert we spoke to, Roth, expressed skepticism: "Do I think that ChatGPT can reliably label really obvious racist speech? Sure, in English. Do I think it can do it in Amharic? Can it do it in nuanced cultural contexts, even in the United States? No, not reliably. I think you need trained humans to do that."[139]

Moreover, online abuse often involves not only text-based content but also images and video, which are more technically difficult to analyze via machine learning. OnlineSOS founder and former head of search and trends policy at Twitter (now X) Liz Lee explained that, compared with text, "detection on photo and video is still lagging."[140] To give just one example, in 2012, Anita Sarkeesian, a feminist media critic, faced a coordinated harassment campaign that included image-based harassment, such as meme templates. She ultimately said that "none of the social media services I use have adequate structures built in to effectively deal with cyber-mob-style harassment."[141]

An Arms Race

Attempting to keep up with the rapidly evolving tactics deployed to hate, harass, and spam is essentially an arms race. Losowsky described it as "always fighting the last war," and offered an illustrative example: "Early on when emojis came out, I don't think they anticipated the eggplant emoji being something that could be abusive. … We, as humans, are just constantly so creative at finding new ways to insult and attack each other."[142] Michelle Ferrier, PhD—past president of the International Association of Women in Radio & Television and founder of TrollBusters, an education organization for journalists facing online abuse—echoed this sentiment: "Our machine learning has not kept up with ways in which our speech, our language, our sentence structures, our visual memes, and other communication culture has changed … and the ways in which people have become very sophisticated about manipulating those digital spaces."[143]

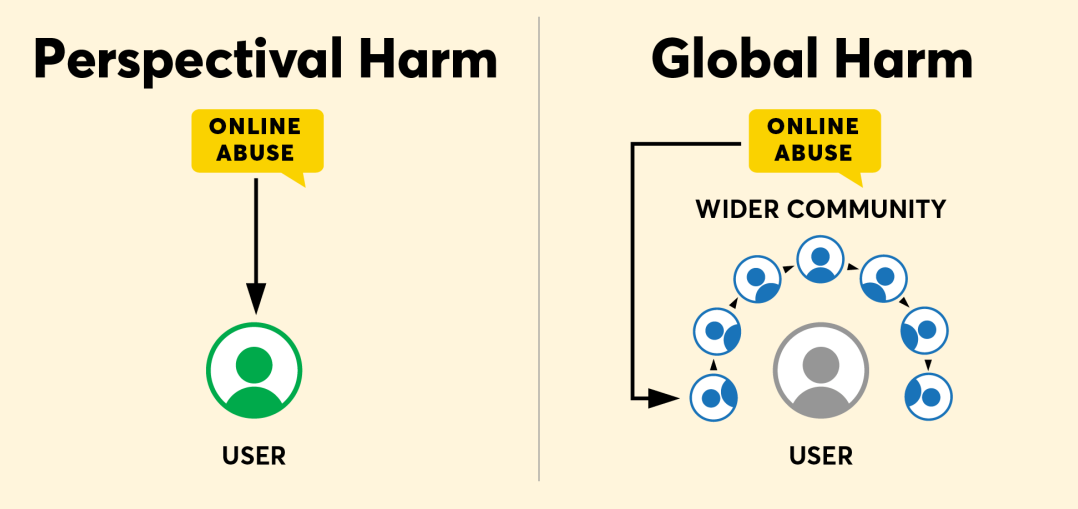
---

[137] Gabriel Nicholas, Aliya Bhatia, "Lost in Translation: Large Language Models in Non-English Content Analysis," *Center for Democracy and Technology,* May 23, 2023, cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/.
[138] Zihan Wang et al., "Extending Multilingual BERT to Low-Resource Languages," *ArXiv,* April 28, 2020, arxiv.org/pdf/2004.13640.
[139] Yoel Roth, interview by Yael Grauer, Consumer Reports, and Deepak Kumar, PEN America and University of California San Diego, December 19, 2023.
[140] Liz Lee, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 7, 2022.
[141] Anita Sarkeesian, "Image Based Harassment and Visual Misogyny," *Feminist Frequency,* July 1, 2012, feministfrequency.com/2012/07/01/image-based-harassment-and-visual-misogyny/.
[142] Andrew Losowsky, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 5, 2022; also described in an email to PEN America from Meta spokesperson, January 2023.
[143] Michelle Ferrier, interview by Yael Grauer, Consumer Reports, January 21, 2022.

While language evolves rapidly, AI models can take considerable time and financial resources to update. By the time a model is ready to deploy, it may already be outdated. The concept is called *model drift*, where a model's predictive power decreases significantly due to underlying shifts in the world around it since it was trained. While it is inherent to the process and inevitable, it can be mitigated only through consistent investment and prioritization.[144]

Perspectival vs. Global Harms

Another distinction between spam and online abuse has to do with whether the harm caused by the problematic content is contingent upon the targeted user actually *seeing* it or whether it can do harm even if the targeted user *never* sees it. Roth calls this distinction a **perspectival** harm vs. a **global** harm.



Graphic demonstrating the difference between perspectival and global harms. Graphic by Consumer Reports.

A perspectival harm is one that is "harmful from the perspective of the viewer." In the case of spam, a user has to see and interact with a misleading piece of spam in order to experience harm by, for example, losing money or having their device compromised. Spam can only cause harm after the user has been exposed to it. Thus, if spam is filtered into a folder, "the harm is effectively neutralized."[145] Indeed, some abusive tactics can also be strictly perspectival, such as a direct personal attack; in those cases, proactive detection and quarantining may be effective at neutralizing the harm.[146] "If you were to call me an idiot," Roth points out, "that would make me feel sad. And if I don't have to see you call me an idiot, then ignorance is bliss."[147]

[144] Jim Holdsworth, Ivan Belcic, Cole Stryker, "What is Model Drift?" *IBM,* July 16, 2024, ibm.com/topics/model-drift.
[145] Yoel Roth, interview by Yael Grauer, Consumer Reports, and Deepak Kumar, PEN America and University of California San Diego, December 19, 2023.
[146] Mike Masnick, "On Social Media Nazi Bars, Tradeoffs, And The Impossibility Of Content Moderation At Scale," *Tech Dirt,* May 4, 2023, techdirt.com/2023/05/04/on-social-media-nazi-bars-tradeoffs-and-the-impossibility-of-content-moderation-at-scale/.
[147] Yoel Roth, interview by Yael Grauer, Consumer Reports, and Deepak Kumar, PEN America and University of California San Diego, December 19, 2023

But some abusive tactics can also cause global harm; they can cause damage regardless of whether the target is exposed or not. For example, doxing, which is the nonconsensual publishing of personally identifiable information, is both a perspectival harm (seeing it can be traumatic and intimidating to the target who is getting doxed) and a global harm (the target does not need to see it to be harmed by it because their private information is now public and available to others to cause additional harm).

Honeywell points out that online abuse can have a different impact depending on whether it is deployed in a private one-on-one environment (email message, DM) or a public environment (in replies or comments on a social media post).[148] When it is public, online abuse can cause reputational damage, which some people simply can't afford not to see. Ferrier says: "You're not seeing it, but that content's gonna be up and online and other people are going to see it, so your reputation is still going to be at risk." She says it can chill the speech of not only the targeted individual but also others who witness it.[149]

Section VI: What Are the Advantages to Treating Online Abuse More Like Spam?

When we talk about platforms treating online abuse more like spam, what we are really talking about is platforms supplementing their own behind-the-scenes content moderation systems with a mechanism that: 1) empowers individual users to have more agency about what potentially abusive content they are exposed to and when, and 2) makes it easier for individual users to take action on abusive content that either violates platform policies or they do not want to see. How this mechanism could work, as described in Section I, is by leveraging automation to proactively detect potentially abusive content (across feeds, threads, comments, replies, direct messages, etc.) and quarantine it in a dashboard, where an individual user could then choose to review and address it with the help of trusted allies—or ignore it altogether.

Platform-driven content moderation systems are critically important for fostering online spaces where users can freely express themselves and exchange ideas and information. Without platform-driven moderation, online spaces would be even more overrun by spam, hate, harassment, threats, disinformation, and other harmful content that can undermine free expression. That is why most technology companies with platforms that facilitate communication among users choose to moderate content to some degree. They draft policies that indicate what content is and is not permitted on their platforms. They rely on a mix of automated systems and professional human moderators to detect and take action on content that violates their policies, which may involve removing the content, downranking it, and/or penalizing the user accounts behind it. In some cases, platforms take action on violative content *proactively,* before most users see it.[150] In other cases, they take action on violative content *reactively,* only after users have seen and reported it.

[148] Leigh Honeywell, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, December 9, 2021.
[149] Michelle Ferrier, interview by Yael Grauer, Consumer Reports, January 21, 2022.
[150] Email to PEN America from Reddit spokesperson, March 2025; email to PEN America from Twitch spokesperson, March 2025; email to PEN America from Meta spokesperson, January 2023.

But platform-driven moderation systems are deeply imperfect for many complex and interrelated reasons. In part, this is because platforms do not sufficiently prioritize and invest resources into content moderation, especially human moderation—as discussed in Section IV. And in part, this is because moderating content at scale is genuinely difficult. "There's this expectation of perfection, that's not actually possible because of the nature of the problem,"[151] Honeywell says. Abusive content, in particular, is often heavily rooted in sociocultural context and individual perception. A significant amount of content falls into legitimate gray areas, where different users may fundamentally disagree about whether it rises to the level of abuse. In a 2021 paper, researchers from the University of Illinois, Stanford University, and Google found that participants disagreed on the toxicity of 85% of the comments they analyzed, highlighting just how challenging the problem of defining online abuse can be.[152]

In attempting to moderate enormous amounts of content globally and at scale, platforms increasingly rely on automation that uses a mix of metadata, heuristics, NLP models, and LLMs. However, as discussed in detail in the previous section, automated systems often get things wrong. Dia Kayyali, independent tech and human rights adviser and former advocacy director at Mnemonic, reinforced this point: "There's all sorts of stuff that's getting taken down incorrectly, and more and more automation is being used."[153] Human moderators, Honeywell says, are "able to understand nuances and situations in a way that automated monitoring cannot yet understand and may never be able to understand," but the problem is scalability. "The downside of human moderation is that all of these social networks are operating at such a scale that it is not actually possible to staff up big enough human moderating teams to deal with the scale that these negative interactions happen."[154]

The consequences of imperfect platform-driven moderation are substantial. On the one hand, platforms may fail to remove content that is abusive or otherwise violates their policies, exposing users to content that can cause substantial harm. On the other hand, they may downrank and remove content that is not abusive or otherwise violative, undermining the free expression of the users who created that content. With that in mind, platform-driven content moderation systems operating behind the scenes are, alone, not enough to sufficiently protect users from online abuse. Such systems make mistakes, they are opaque, they can be blunt in that they often impact all users across the board, and they centralize control and power in the hands of private corporations and their top leadership. Instead, we propose that platforms complement their behind-the-scenes moderation systems with a user-driven mechanism like the one explored in this report. Below we outline the advantages of such an approach.

Protects Mental Health by Reducing Exposure to Potential Abuse

Most existing in-platform features to address online abuse, such as blocking, muting, or reporting, are reactive: a user must first see and experience the abuse before they can take action. The problem with reactive measures is that they require a user to be exposed to abusive content, often repeatedly, which can negatively impact mental health, sometimes severely. As outlined in the introduction, online abuse can lead to severe mental health issues—such as

[151] Leigh Honeywell, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, December 9, 2021.
[152] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey, "Designing toxic content classification for a diversity of perspectives," *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*, SOUPS'21, Article 16, 299–317.
[153] Dia Kayyali, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 7, 2022.
[154] Leigh Honeywell, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, December 9, 2021.

depression, anxiety, post-traumatic stress disorder, and in extreme cases, self-harm or suicide—and it can push people offline and even out of their professions.[155]

As things stand now, users may be confronted with abusive content with little to no warning whenever they access a digital communications platform, whether via their social media feeds, DM's, or emails. The uncertainty around whether or not they will encounter hate and abuse only amplifies its effects, which can be damaging. As a clinical and research neuroscientist explained in PEN America's 2024 report "The Power of Peer Support": "Ambiguity and uncertainty is very expensive, metabolically. An uncertain outcome—which might be either really bad or might be OK—is actually [physiologically] more costly than knowing 100 percent [that] something bad is going to happen."[156]

The mechanism we propose here ensures that a user does not have to experience that content by default. Instead, users have the choice to review the content, if and when they are ready, or ignore it altogether. To further mitigate damage to mental health, the mechanism could be designed with trauma-informed principles, clearly flag potentially dangerous content to facilitate monitoring for threats, and enable users to delegate review to a trusted ally so that they do not have to be exposed to the abusive content.[157] We explore all three of these recommendations in more detail in Section VII.

Protects Free Expression

One of the biggest challenges with platform-driven content moderation is the balance between protecting users from potentially harmful content with protecting users' freedom of expression. When platforms tighten their automated detection systems behind-the-scenes to more aggressively filter out potentially abusive content, they risk mistakenly sweeping up more content that was not intended or experienced as abusive and restricting its visibility for all users. When an individual user tightens the automated detection of potentially abusive content using a mechanism like the one we propose, they are restricting its visibility only for themselves. Each user can then review whatever content was filtered out and adjust the automated filtering mechanism until they find the balance right for them. Because the mechanism proposed here is focused on empowering individual users to fine-tune what they themselves see and when—rather than controlling what all other users can see and access—it allows for a much broader range of content. In other words, users can better protect themselves from online abuse without infringing on the free expression of others.

Empowers Regular Users

Platforms are already using increasingly powerful automated systems to proactively detect potentially abusive content, but these systems are largely available only to their own staff and

[155] Julie Posetti, Nabeelah Shabbir et al., "The Chilling: global trends in online violence against women journalists," *UNESCO,* April 12, 2021, unesdoc.unesco.org/ark:/48223/pf0000377223; Jackson Richman, "Taylor Lorenz Breaks Down on MSNBC Sharing Experience Being Targeted Online, Contemplated Suicide," *Mediaite,* April 1, 2022, mediaite.com/tv/taylor-lorenz-breaks-down-on-msnbc-sharing-experience-being-targeted-online-contemplated-suicide/; Ben Dooley and Hikari Hida, "After Reality Star's Death, Japan Vows to Rip the Mask Off Online Hate," *The New York Times,* June 1, 2020, nytimes.com/2020/06/01/business/hana-kimura-terrace-house.html.

[156] "The Power of Peer Support," *PEN America,* September 25, 2024, pen.org/report/peer-support.

[157] Catherine Han, Anne Li, Deepak Kumar, and Zakir Durumeric, "PressProtect: Helping Journalists Navigate Social Media in the Face of Online Harassment," *Computers and Society (cs.CY) Human-Computer Interaction (cs.HC)*, January 2024, arxiv.org/abs/2401.11032.

third-party moderation companies rather than to individual users. What we're proposing is that platforms empower regular users to leverage some of these sophisticated automated detection mechanisms to mitigate their own exposure to hateful or harassing content. Within the centralized dashboard of the mechanism proposed here, users would be able to review, take action on, and adjust the sensitivity of the filters for: 1) content that violates platform policies but was not appropriately addressed by platform-driven moderation systems, 2) content that the user believes violates platform policies but that the platform has determined does not violate its policies, and 3) content that does not violate platform policies but that the user still wants to limit their exposure to. This would give users greater agency and control to adjust the degree and volume of all kinds of abusive content they encounter on their own feeds, DMs, etc., which multiple recent studies have shown that many users appreciate and want. In 2022, researchers from Rutgers University and the University of Washington created FilterBuddy, a tool that enables YouTube content creators to have more control over their audience engagement by allowing them to create word filters, which are lists of blocked terms, for their comment sections. The researchers found that creators greatly appreciated the additional control enabled by the word filters, and ultimately saw it as a promising, effective way to moderate their channel.[158] A 2023 study further found that 52% of participants preferred personal content moderation of potential abuse and harassment over platform-wide controls.[159] Another 2023 study found that social media platform users generally desired greater control over personal moderation, with a particular desire for greater transparency and knowledge of how moderation systems were functioning.[160]

Increases Transparency

Although most platforms have publicly visible policies, the content moderation systems they use to implement those policies operate behind the scenes and are largely opaque. When platforms proactively detect violative content and remove or downrank it, most users never see that content. Users have to deal with the consequences of content moderation with little understanding of how, why, and what content never reaches them.[161] A lack of transparency around what speech is moderated and why can lead to unequal censorship, especially for communities whose voices are already underrepresented or marginalized. What we propose in this report is a more sophisticated automated filtering mechanism that allows users to see—in a centralized interface—all of the content that has proactively been detected, filtered out, and quarantined. While seeing quarantined content would be purely optional, giving users and their allies the choice to easily review it in one place would make automated filtering more transparent. Combined with the ability to fine-tune automated filtering, users could make adjustments until they found a personalized threshold that works best for them. The importance of including a centralized user interface, like a dashboard, and enabling user fine-tuning of automated filters are among the recommendations explored in more detail in Section VII.

---

[158] Shagun Jhaver, Quanze Chen, Detlef Knauss, and Amy Zhang, "Designing Word Filter Tools for Creator-led Comment Moderation," *Association for Computing Machinery*, April 29, 2022, Article no. 205, 1–21. doi.org/10.1145/3491102.3517505.
[159] Shagun Jhaver and Amy Zhang, "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 and Society* (2023), doi.org/10.1177/14614448231217993.
[160] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang, "Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor," *ACM on Human-Computer Interaction* 7, no. 289 (October 2023), 1-33, dl.acm.org/doi/pdf/10.1145/3610080.
[161] Sarah Roberts, "The Great A.I. Beta Test," *Slate*, April 8, 2020, slate.com/technology/2020/04/coronavirus-facebook-content-moderation-automated.html.

Has Potential to Improve Platform-Driven Content Moderation

Depending on how it is designed and integrated, the user-facing mechanism we propose here has the potential to gradually strengthen behind-the-scenes platform-driven content moderation. The mechanism asserts the importance of keeping a human in the loop—specifically the human impacted by the potentially abusive content and/or their trusted allies. Such a mechanism makes it easier for regular users to review and take action on potentially abusive content because it uses powerful automated systems to proactively detect and quarantine all of that content in one place, where users can also easily access all of the functionalities to take action on that content. Action may involve reporting the content to the platform if the user feels it violates platform policies, blocking or muting the accounts behind the content, or removing the content from quarantine because the user determines that it is a false positive or does not perceive it as abusive or harmful. Such user-driven actions can and should serve as useful ongoing signals to refine and improve behind-the-scenes platform-driven moderation. Empowering users with a mechanism to feed their own experiences of abuse into platform-driven content moderation systems also provides platforms with crucial training data on how users perceive content according to their individual preference and tolerance levels.

Section VII: Recommendations

If technology companies are going to give regular users access to a powerful mechanism that proactively detects abuse, the design of that mechanism—particularly the user interface—is critically important. Based on our desk research, interviews, experience, and expertise, we lay out a set of recommendations for how to build an effective mechanism that automatically detects and quarantines abusive content on online communications platforms in a way that is user-friendly, protective of a user's mental health, and protective of free speech.

Create a Centralized Dashboard for Review and Action

Several of the people we interviewed stressed the importance of any automated filtering mechanism having a user-friendly interface—essentially a dashboard—that centralizes: 1) all content that has been detected and quarantined for review, 2) all actions users can take on that content, and 3) all fine-tuning functionalities.[162] Sinders pointed out that: "Even if you're just getting regular spam, regular weird messages, and you have a lot of replies, having a dashboard to sort them is super helpful."[163] By putting all features to address abuse in one place, platforms can save users time and energy from having to dig through multiple sections of an app or a website to understand and leverage their options to protect themselves.

Creating a centralized place where users can review and take action on abusive content is critical not only for convenience and efficiency but also for the mitigation of psychological harm. When a user accesses their regular feed, DMs, or any other surface within a social media or other communications platform, being immediately confronted with abuse and threats can be traumatizing. If the potentially abusive content is proactively detected and hidden from the main or default platform surfaces—and quarantined in a dashboard—then users can make a deliberate, conscious decision to access that dashboard if and when they are ready to navigate potentially abusive content.

[162] Liz Lee, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 7, 2022.
[163] Caroline Sinders, interview by Yael Grauer, Consumer Reports, December 14, 2021.

Existing Example: Twitch offers a customizable dashboard called Mod View, which allows user-moderators to centralize and organize all of the tools they need to moderate platform channels. Within Mod View, in the AutoMod Queue, Twitch centralizes all messages that have been quarantined via the AutoMod feature so that moderators can easily review and approve or reject these messages.[164]

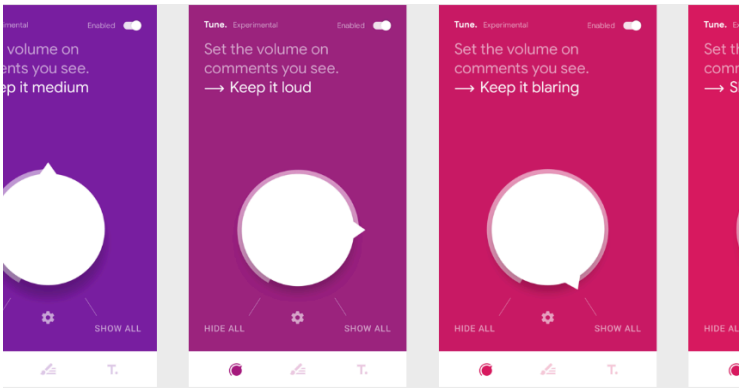
Allow User Fine-Tuning

Technology companies must allow for user fine-tuning and flexibility in an automated filtering mechanism. No two users are alike, and every user will have their own preferences for what content they would like to see and engage with and what content they would want to be detected and quarantined from their feeds, DMs, inboxes, etc. Some users will prefer not to filter out anything, while others may prefer to filter out the majority of abusive content. Journalists, for example, may simultaneously want to protect themselves from harassment while still needing to be aware of how a particular story they have published is being received. "As a class of citizens and professionals whose responsibility it is to put out truthful information and to be in these public spaces … we need to be able to see these attacks,"[165] Ferrier said. In other words, every user, depending on their personal and professional circumstances, will have a different threshold for what potentially abusive content should be quarantined.

Enabling user fine-tuning addresses a concern commonly expressed by the people we interviewed: the risk of an overactive abuse filter, which may make users less likely to take advantage of such a mechanism. "My expectation would be there will be a fairly low tolerance for false positives," Edelson said. "People would be unhappy if a message from their friend was put in their toxic [harassment] folder when it was not."[166] A Meta spokesperson also pointed out that a casual or social user of Instagram might have a higher tolerance level than, for example, a business account manager who might miss opportunities if they are flagged as inappropriate or potentially abusive.[167] Such a mechanism will be significantly more useful if users are able to fine-tune the threshold for abusive content, or its level of sensitivity, to suit their individual needs.

Existing Example: Tune, an experimental web browser extension launched by Google's Jigsaw team in 2019, provides a promising example. Tune leverages machine learning to enable users to fine-tune the level of toxicity they see across digital platforms.[168]

[164] "Guide for Moderators," *Twitch,* accessed March 11, 2025, https://safety.twitch.tv/s/article/Guide-for-Moderators?language=en_US.
[165] Michelle Ferrier, interview by Yael Grauer, Consumer Reports, January 21, 2022.
[166] Laura Edelson, interview by Deepak Kumar, PEN America and University of California San Diego, and Viktorya Vilk, PEN America, May 21, 2024; also mentioned in Caroline Sinders, interview by Yael Grauer, Consumer Reports, December 14, 2021.
[167] Email to PEN America from Meta spokesperson, January 2023.
[168] CJ Adams, "Tune: Control the comments you see," *Medium,* March 12, 2019, medium.com/jigsaw/tune-control-the-comments-you-see-b10cc807a171.



Screenshot of the adjustable anti-harassment filters on Google's Tune extension. Photo by BetaNews.

Prioritize Automated Learning and Personalization

A sophisticated automated filtering mechanism should gradually be able to learn from individual user needs, preferences, and behavior in order to improve its ability to accurately gauge what content that user will want to see and engage with vs. what content they will want filtered out and quarantined for later review. In the case of automated abuse filters, according to Losowsky, "you can start to train a model based on your own preferences. You can improve the model based on what you … approve, reject, block, and so on. … Personalized models that learn over time are a major aspect of computer-assisted moderation."[169] A trained model would, over time, allow for more precision and personalization. To reduce the burden of training a model from scratch, platforms could introduce crowdsourcing, where users could import the preferences and filters of their friends, allies, or colleagues, and adjust accordingly from there. This is not without precedent. Many email spam filters do this already; for example, if an individual marks a message as spam, future messages with those same characteristics will also be filtered as spam in that individual's inbox.

Creating a mechanism that not only can learn but also is personalized to each user, is a substantial undertaking that would require significant investment and computing power. Edelson expressed skepticism that platforms would have the financial incentives to build personalized toxicity detectors, but said: "If anyone has the context to do this, it would be social media platforms, because a lot of this is going to be context specific to the sender and the recipient … remember that they build personalized algorithms all the time."[170] In other words, because platforms have a long history of building automated learning and personalization algorithms to grow revenue, they already have the experience and expertise to create an individualized anti-abuse mechanism that improves over time.

[169] Andrew Losowsky, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 5, 2022.
[170] Laura Edelson, interview by Deepak Kumar, PEN America and University of California San Diego, and Viktorya Vilk, PEN America, May 21, 2024.

Existing Example: On Bluesky, users can plug in Ozone, an open-source tool that enables users to build their own highly customized automated filters and share them with other users, facilitating the kind of crowdsourcing recommended above.[171]

Flag Particularly Dangerous Content

Any automated filtering mechanism will need to balance reducing exposure to abuse against inadvertently masking especially egregious tactics that require assessment and, potentially, action. Multiple people we interviewed stressed that it was critical for such a system to automatically detect and flag potentially dangerous content, such as death threats, threats of sexual or physical violence, and doxing.[172] Hiding or quarantining abusive content "helps to dampen your exposure," according to Ferrier, but "it can heighten your fear and your risk and your safety because you don't know what's going on."[173] Expressing a similar concern, Kayyali said: "If something just gets hidden from you, you have nowhere to check what's been hidden from you, or what's been removed from your field of vision that might actually put you at risk." To address this, Kayyali suggested that death threats, "instances of someone's address or phone number being posted," and other potentially dangerous tactics be proactively tagged, potentially with readily visible "graphical little red flag."[174]

Several of the sources we interviewed for this report, including Roth and Edelson, also argued that automatically detecting and flagging the most egregious abusive tactics is more technically feasible than detecting abusive content that falls into a gray area. "It's pretty tractable to build a death threats classifier. … It's pretty tractable to build a classifier [for] explicit acts of violence," Edelson said, as well as "nudes detector," but "as you walk down the path from that, it's going to get harder and harder."[175]

Leverage Trauma-Informed Design

One of the primary purposes of the mechanism proposed here is to alleviate psychological harm by reducing exposure to hate and harassment. However, because many users may need or choose to review abusive content that has been quarantined, some degree of exposure may be necessary. Having to scroll through a dashboard of abusive content to find high-quality messages that have been misclassified as abusive can cause emotional distress. "I remember early spam folders, you would end up calling people and being like, 'go into your spam folder,'

[171] The Bluesky Team, "Bluesky's Stackable Approach to Moderation," *Bluesky,* March 12, 2024, bsky.social/about/blog/03-12-2024-stackable-moderation.
[172]Kat Lo, interview by Deepak Kumar, PEN America and University of California San Diego, and Viktorya Vilk, PEN America, May 21, 2024; Yoel Roth, interview by Yael Grauer, Consumer Reports, and Deepak Kumar, PEN America and University of California San Diego, December 19, 2023; Laura Edelson, interview by Deepak Kumar, PEN America and University of California San Diego, and Viktorya Vilk, PEN America, May 21, 2024; Dia Kayyali, interview by Yael Grauer, Consumer Reports and Viktorya Vilk, PEN America, January 7, 2022; Liz Lee, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 7, 2022.
[173] Michelle Ferrier, interview by Yael Grauer, Consumer Reports, January 21, 2022.
[174] Dia Kayyali, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 7, 2022.
[175] Laura Edelson, interview by Deepak Kumar, PEN America and University of California San Diego, and Viktorya Vilk, PEN America, May 21, 2024; Yoel Roth, interview by Yael Grauer, Consumer Reports, and Deepak Kumar, PEN America and University of California San Diego, December 19, 2023.

and almost every day you would have to go into a spam folder and take a look and wade through that," Losowsky said. "Now that would be fine, because it was just wading through fake Viagra ads. If you had to wade through personal abuse, it's just a very different emotional toll that is put on you."[176]

For these reasons, it is imperative that the dashboard for such a mechanism be designed with trauma-informed principles in mind, such those developed by the Centers for Disease Control and Prevention and the Substance Abuse and Mental Health Services Administration. These principles include safety, trustworthiness, transparency, peer support, collaboration, mutuality, empowerment, voice, choice, and sensitivity to cultural, historic, and gender issues.[177] Scholars elsewhere have explored what these pillars can look like from a design perspective. Such a dashboard could, for example, blur all content by default and users could then click to unblur it, either for each individual piece of quarantined content or as a setting across the board. It could be transparent about how content is assessed and flagged to its user in order to increase trustworthiness. Relaxing colors and simple design, self-care nudges when users are viewing difficult content, quick exit options, and thematic trigger warnings (for example, if the content involves mentions of sexual violence, racial slurs, etc.) are other design features that can foster a more trauma-informed dashboard.[178]

Facilitate Delegation

The mechanism proposed in this report should enable users to designate a limited number of trusted allies to help them take action on the abusive content that has been detected and quarantined in their dashboard. This can be done through a "delegated access" system, akin to that available on Gmail. These delegates could assist with tasks such as blocking, muting, reporting, and documenting abuse. Users should be able to control the level of access and control their delegates have.

This is important because reviewing and taking action on potentially abusive content, even when it has been proactively quarantined to a centralized dashboard, can be overwhelming and exhausting. For people who have experienced severe trauma in non-online contexts, strong social support has been shown to significantly reduce harm and build resilience.[179] Baking social support into a mechanism designed to mitigate online abuse may lead to similar improvements in psychosocial wellbeing for its targets.

Existing Example: Delegated access exists not only in Gmail but also existed in at least one anti-harassment tool—Block Party, which we discuss in detail in Section II. The tool had a Helper View function, which allowed users to delegate blocking and muting features to a trusted

[176] Andrew Losowsky, interview by Yael Grauer, Consumer Reports, and Viktorya Vilk, PEN America, January 5, 2022.
[177] "6 Guiding Principles to a Trauma Informed Approach Infographic," *Centers for Disease Control and Prevention,* 2022, stacks.cdc.gov/view/cdc/138924.
[178] Carol F. Scott, Gabriela Marcu, Riana Elyse Anderson, Mark W. Newman, and Sarita Schoenebeck, "Trauma-Informed Social Media: Towards Solutions for Reducing and Healing Online Harm," *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23) (New York: Association for Computing Machinery, 2023),* Article 341, 1–20, doi.org/10.1145/3544548.3581512.
[179] Wenjie Dai et al., "Association between Social Support and Recovery from Post-Traumatic Stress Disorder after Flood: A 13–14 Year Follow-Up Study in Hunan, China," *BMC Public Health* 16 (2016): 194, doi.org/10.1186/s12889-016-2871-x.

ally without granting them full password permission to their impacted Twitter (now X) account.[180] "The biggest feature of Block Party that was useful was the collaborative component," said Kat Lo, content moderation lead at internet equity nonprofit Meedan. "So if somebody else being able to access [a spam-like abuse] folder is straightforward or easy, then it's especially valuable."[181]

Facilitate Documentation

A mechanism that is already automatically detecting online abuse and quarantining it is uniquely well positioned to facilitate documentation. All users facing abuse need to have the ability to maintain a detailed, exportable record of what they have been subjected to. Documentation can help targeted users track online abuse and facilitate communication with allies, employers, and lawyers. In a paper published by Google, researchers showed that documentation helps those targeted establish proof that their experience is real, and that this proof is often necessary to escalate the abuse to platforms or law enforcement.[182] The pressing need for a documentation feature is underscored by the fact that targets can lose evidence if, for instance, an abuser deliberately deletes the content as soon as it has been seen or if the content is reported, determined to be a violation of platform policies, and removed. As such, even if abusive content is deleted by the user who created it, a copy should be maintained in the quarantine area for the targeted user.

Existing Example: The TRFilter tool initially developed by Google's Jigsaw team and launched by Thomson Reuters Foundation—which is discussed in Section II—was explicitly designed to leverage machine learning for automated detection in order to facilitate documentation of abuse.

Leverage Multiple Detection Signals to Address Coordinated Harassment

A sophisticated mechanism will also combine detection of multiple different signals, which would include the use of machine learning and, potentially, eventually LLMs to analyze content, as well as the analysis of metadata about the content and user behavioral patterns and other heuristics. To detect abuse effectively, Edelson says, mechanisms "need to rely on some knowledge of the messaging behavior of the senders, their history of messages between the sender and the recipient."[183] Signals could include whether or not an account was created recently, whether the account looks automated (for example, the username has a random string of numbers or has no

[180] "How it works: Helper view," *Block Party,* accessed September 12, 2024, blockpartyapp.com/how-it-works/#:~:text=Get%20support%20from%20your%20friends,access%20to%20your%20Twitter%20account.
[181] Kat Lo, interview by Deepak Kumar, PEN America and University of California San Diego, and Viktorya Vilk, PEN America, May 21, 2024.
[182] Nitesh Goyal, Leslie Park, and Lucy Vasserman, "'You have to prove the threat is real': Understanding the needs of Female Journalists and Activists to Document and Report Online Harassment," *2022 CHI Conference on Human Factors in Computing Systems*, no. 242 (April 2022), 1-17, dl.acm.org/doi/pdf/10.1145/3491102.3517517.
[183] Laura Edelson, interview by Deepak Kumar, PEN America and University of California San Diego, and Viktorya Vilk, PEN America, May 21, 2024.

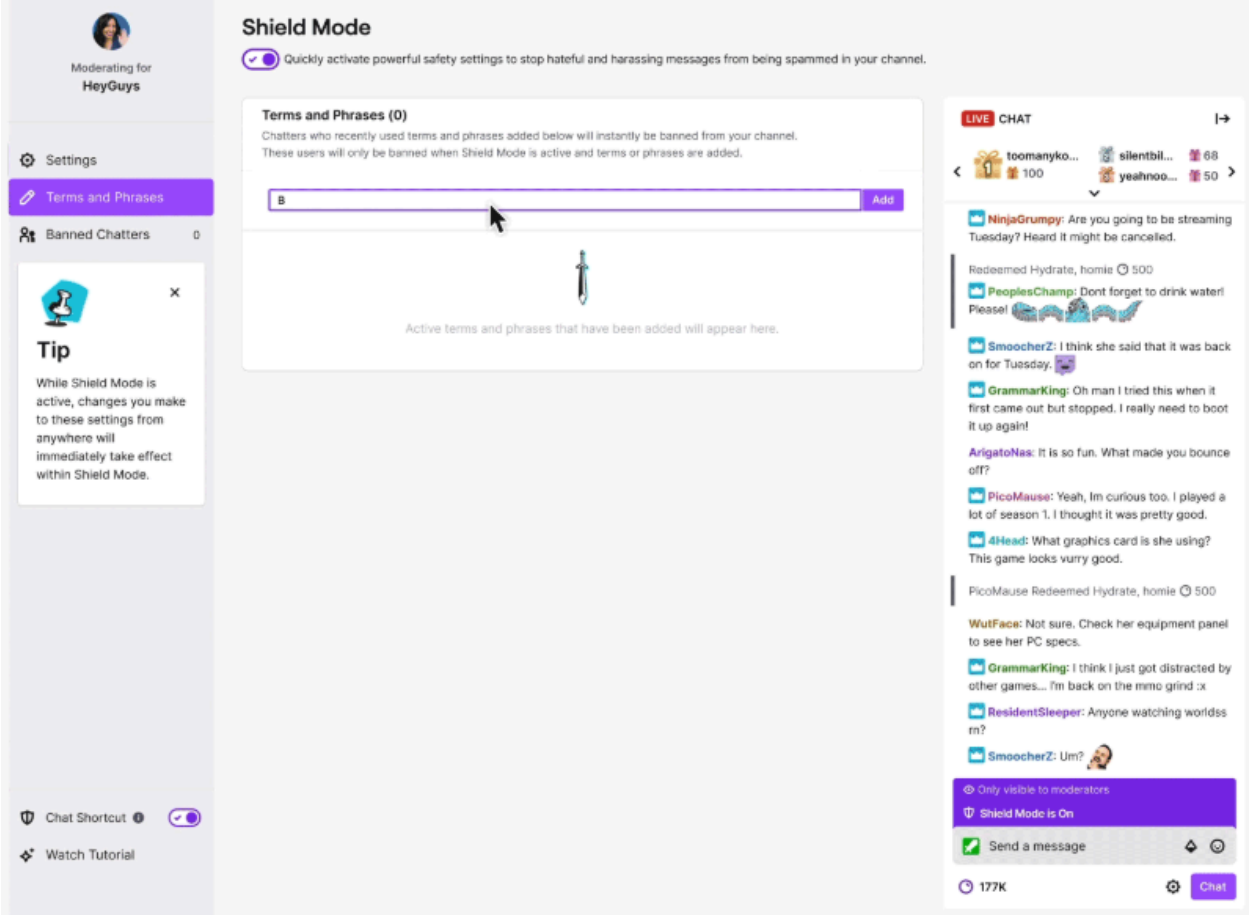
image), whether an account has shared connections with the account it is attempting to interact with (followers, friends, etc.), whether an account has been blocked repeatedly by other accounts, whether the account sends out a lot of content at once, and whether there are multiple accounts created from a single IP address.

One significant advantage of thinking about abuse more like spam is that spam is often conceptualized and tackled at scale, and there is recognition that it can be coordinated and networked, whereas abuse is often still addressed piecemeal. In fact, research points to abuse being increasingly coordinated and orchestrated. Loose networks seeded on platforms like 4chan and Kiwi Farms have coordinated "pile on" campaigns such as Gamergate's attack on women video game designers,[184] and the severe targeting of women journalists like Taylor Lorenz, Carole Cadwalladr, and Maria Ressa,[185] and women politicians like Alexandria Ocasio-Cortez.[186] Michelle Ferrier of TrollBusters calls such coordinated campaigns "smart mobs," which she describes as "coordinated activities by groups like white supremacists, domestic terrorists … bad political actors who use coordinated campaigns and multiple attacks across platforms."[187]

A lack of understanding or interest in the degree of coordination behind harassment campaigns can lead to inconsistent, patchy responses from platforms. Describing a period when multiple prominent Black women were so severely harassed on Twitter (now X) that they quit the platform, Anika Navaroli, a former senior policy official at Twitter (now X), noted, "we weren't necessarily taking into account [these larger networks] because [we were] very much assessing content piece by piece vs. the larger approach of seeing this as spam."[188] As a result, content moderation was more cumbersome and less effective.

Existing Example: Twitch's Shield Mode allows users (both streamers and moderators) to "turbocharge" safety settings with one click; the feature was designed, in part, to respond to hate raids and coordinated harassment campaigns, during which users often need to activate stronger protections easily and on short notice.[189]

[184] St. James, Emily. "#Gamergate: Here's Why Everybody in the Video Game World Is Fighting." *Vox*, October 13, 2014, vox.com/2014/9/6/6111065/gamergate-explained-everybody-fighting.
[185] Julie Posetti and Nabeelah Shabbir, "The Chilling: A Global Study On Online Violence Against Women Journalists," *International Center for Journalists,* November 2, 2022, icfj.org/our-work/chilling-global-study-online-violence-against-women-journalists.
[186] "DISINFORMED: Alexandria Ocasio-Cortez Is Facing a Coordinated Disinfo Attack | There Are No Girls on the Internet Podcast," *Everand,* February 6, 2021, everand.com/podcast/591378144/DISINFORMED-Alexandria-Ocasio-Cortez-is-facing-a-coordinated-disinfo-attack.
[187] Michelle Ferrier, interview by Yael Grauer, Consumer Reports, January 21, 2022; Michelle Ferrier and Nisha Garud-Patkar, "TrollBusters: Fighting Online Harassment of Women Journalists," *Mediating Misogyny* (2018): 311-332. https://doi.org/10.1007/978-3-319-72917-6_16.
[188] Liz Lee and Anika Navaroli, interview by Deepak Kumar, PEN America and University of California San Diego, May 7, 2024.
[189] "Protect Your Channel With Shield Mode," *Twitch,* accessed March 11, 2025, https://safety.twitch.tv/s/article/Protect-your-channel-with-Shield-Mode?language=en_US.



Screenshot of Twitch's Shield Mode tool. Photo by Twitch.

Include Impacted Stakeholders From the Beginning

Any mechanism intended to proactively and robustly protect users from abuse must be developed, from its earliest stages through to its completion: 1) by a diverse team of product managers and engineers, in terms of identity and lived experience, and 2) in consultation with people disproportionately targeted for hate and harassment. Several of the people we interviewed who had worked in technology companies forcefully underscored both points. Navaroli, addressing technologists who are building anti-harassment tools, asserts: "Talk to those of us who have done this work … who hold these identities. … If we're thinking about abuse… who are these things being directed towards? Let's have those folks in the process of creating the solutions."[190] As Lee pointed out, it is especially important that those with the most influence in a technology company, the decision-makers, have insight into how online abuse unfolds and whom it most impacts: "It's not just about the team's racial composition. It's about people understanding communities who may be under threat or impacted differently. And it's not

[190] Liz Lee and Anika Navaroli, interview by Deepak Kumar, PEN America and University of California San Diego, May 7, 2024.

even about the people who are working on the topic. It's more the people who are making decisions, do they understand the communities on their platforms?"[191] Given that the mechanism discussed in this report will rely heavily on machine learning—especially in light of the challenges with implicit bias discussed in Section V—technologists will need to diversify the sources of their data sets, audit and counterbalance inherent biases in the data sets, and build equitable pipelines in the annotation of data sets to ensure that AI models will not inadvertently further harm marginalized communities.

Complement, Rather Than Replace, Platform-Driven Content Moderation

Finally, platforms should approach user-driven mechanisms like the one we propose here as a complement to—rather than as a replacement for—their own behind-the-scenes content moderation systems. To effectively protect users from online abuse and other harmful content, platforms must invest in both. The mechanism we propose here is best suited to alleviate perspectival rather than global harms. As explained in detail in Section V, perspectival harms are damaging from the perspective of the targeted user and predicated on their exposure to the content; global harms, on the other hand, can cause damage to the targeted user and/or to a wider community—even if the targeted user never sees the content. To give one example, a social media post that includes a user's home address can endanger the targeted user whether they see the post or not because other users now know where the targeted user lives. In a case like that, given that doxing violates most platforms' policies, it is not enough for a platform to quarantine the abusive content so that the targeted user does not have to see it; the platform needs to remove that content altogether as part of platform-driven content moderation systems that enact platform policies. To give another example, if a prominent person of color is bombarded with racial slurs and violent threats whenever they post on a social media platform, the abusive tactics risk chilling not only the targeted user's speech but also the speech of other users with a shared identity because they, too, could face threats and slurs whenever they speak. As Roth, Twitter's former head of Trust and Safety, points out, "a purely victim-focused spam folder approach doesn't protect the community from abuse."[192]

In other words, when it comes to abusive tactics likely to cause global harm—such as violent threats and doxing of sensitive private information—platforms must continue to invest in and improve their own behind-the-scenes content moderation efforts to enforce their own policies and protect their users. And while platform-driven content moderation will, in all likelihood, increasingly rely on automation, platforms must continue to invest in professional human moderators as well. As Theodora Skeadas, former associate on public policy at Twitter (now X), explains, "Technology solutions alone are typically inadequate. You need a human in the loop … because the technical systems are themselves imperfect."[193] The good news is that user-driven mechanisms—as explained in Section VI—can actually substantially strengthen platform-driven content moderation systems.

[191] Liz Lee and Anika Navaroli, interview by Deepak Kumar, PEN America and University of California San Diego, May 7, 2024.
[192] Yoel Roth, interview by Yael Grauer, Consumer Reports, and Deepak Kumar, PEN America and University of California San Diego, December 19, 2023.
[193] Theodora Skeadas, interview by Deepak Kumar, PEN America and University of California San Diego, January 5, 2022.

Conclusion

If the technology companies that run digital communications platforms, including social media and email, are serious about protecting free speech, they must do more to protect their users from abuse. This means that platforms must invest in and strengthen their behind-the-scenes content moderation systems to ensure that they are upholding their own hate, harassment, and cyberbullying policies more fairly, accurately, efficiently, and transparently. However, platform-driven moderation alone—given the challenges inherent to adjudicating online abuse while protecting free speech—is not enough.

As we have argued here, technology companies must also invest in features that empower individual users to shape their own experiences on digital platforms. Existing user-driven features, such as blocking and reporting, are critically important. But they are also insufficient because they are *reactive;* in other words, they require users to be repeatedly exposed to abusive content before it can be addressed, which can detrimentally impact targets' mental health and lead to self-censorship.

In this report, we propose one particularly promising mechanism that can proactively protect users from online abuse: treating online abuse more like spam by detecting and quarantining it for users to review and address—if and when they are ready to do so. In this report, we map out the challenges and opportunities of developing such a mechanism and outline key recommendations to ensure that it is user-friendly, effective, and responsive. We come to the conclusion that treating online abuse more like spam, while it may pose unique technological challenges, is worthwhile because it has numerous advantages: It can provide users more agency over their online spaces, increase transparency and trust between users and platforms, improve platform-driven content moderation, and reduce the psychological toll of online harassment while protecting users' free speech.

We recognize that this proposal, which is essentially technological, is only part of the puzzle when it comes to tackling online abuse. Other important factors, including regulation, platform governance, and shifting social norms, all play a crucial role. While examining these factors is beyond the scope of this report, we acknowledge that technical solutions alone can never be a panacea to reducing online harms. We hope that by mapping out one specific, concrete, actionable solution to empower and protect users in the face of online abuse, our research presents a constructive path forward for technologists, policymakers, civil society, and platforms. We call on technology companies, particularly social media platforms, to rise to the challenge of developing proactive protections against online abuse that are more robust, user-friendly, and trauma-informed, like the one we propose here, in order to build spaces for public discourse online that are safer, more equitable, and more free.

Methodology

Collectively, the authors of this report have extensive experience in providing support for people facing online hate, harassment, and other forms of intimidation and in studying and writing about technology and online harms for the media, academia, and civil society. The authors—sometimes together and sometimes separately—conducted 12 semi-structured interviews, over two years, with 14 subjects who worked in a variety of relevant contexts, ranging from civil society organizations and universities to startups and massive global tech companies—including as trust and safety professionals. Most interviews were on the record, though several were on background. The majority of people we interviewed identify as women or nonbinary and as people of color.

Interviews were augmented with extensive desk research—with sources ranging from journalism to published academic papers from the computing and communications sectors. We reached out to the following platforms for formal commentary: Bluesky, Google, Meta, Reddit, TikTok, Twitch, and X; we heard back from and incorporated input from Meta, Reddit, and Twitch. And we also reached out for fact-checking purposes to the following companies that have developed third-party anti-harassment tools, all of which responded: BlockParty, Coral, Jigsaw, Squadbox, and TRFilter. Finally, four expert reviewers—with backgrounds studying computer science, developing technology products, investigating online abuse, and working in trust and safety at technology companies—provided detailed external feedback on the report.

Online abuse is a phenomenon with many names—online harassment, cyberbullying, online violence, technology facilitated gender-based violence, etc.—and no universally agreed-upon definition. Platforms use a variety of different definitions and terms in their policies—primarily "harassment," "cyberbullying," "bullying," "hateful conduct," and "hate speech." In this report, we use the terms "online abuse" and "online harassment" interchangeably; PEN America defines these terms as the "pervasive or severe targeting of an individual or group online through harmful behavior."[194] We use alternative terms only in cases where we are discussing a platform's specific policies, in which case we use the terminology used by that platform in that policy.

[194] "Defining 'Online Abuse': A Glossary of Terms," Online Harassment Field Manual, *PEN America,* onlineharassmentfieldmanual.pen.org/defining-online-harassment-a-glossary-of-terms/.

Acknowledgments

This report was co-authored by Yael Grauer, program manager of cybersecurity research at Consumer Reports; Deepak Kumar, assistant professor of computer science and engineering at the University of California San Diego; and Viktorya Vilk, director of digital safety and free expression at PEN America. Amanda Wells, program assistant for digital safety and free expression at PEN America, conducted additional writing, editing, and research. It was designed by Chris Griggs, associate director, brand creative at Consumer Reports, and copyedited by Wendy Greenfield, copy chief at Consumer Reports. PEN America's Interim co-CEO and chief program officer of free expression programs, Summer Lopez, reviewed the report and offered thoughtful feedback, as did James Tager, PEN America's director of research, and Hanna Khosravi, PEN America's program manager of research. We would also like to thank the current and former PEN America program assistants and interns whose research, fact-checking, and proofreading made this report possible: Doris Zhang, Aashna Agarwal, Zoe Briscoe, Luke Flyer, Wyatt King, and Victoria He.

PEN America and Consumer Reports extend special thanks to the following experts for providing invaluable input on this report: Morgan Sung, journalist and host of "Close All Tabs" from KQED in San Francisco; Allison McDonald, assistant professor of computing and data sciences at Boston University; Amy Zhang, assistant professor at the University of Washington Paul G. Allen School of Computer Science & Engineering; and Azmina Dhrodia, policy expert on tech-facilitated gender-based violence and former safety policy lead at Bumble. We are deeply grateful to the many journalists, writers, scholars, technologists, civil society advocates, and other experts who agreed to be interviewed for this report, including those who are not acknowledged by name. PEN America and Consumer Reports appreciate the responsiveness of the representatives at Meta, Reddit, and Twitch in our exchanges.

Our deep abiding appreciation goes to Craig Newmark Philanthropies for supporting this project.

PEN America has previously received financial support from Google and Meta, but those funds did not support the research, writing, or publication of this report.