NO EXCUSE FOR ABUSE

What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users

EXECUTIVE SUMMARY

INTRODUCTION

Online abuse—from violent threats and hateful slurs to sexual harassment and impersonation—is a pervasive and growing problem.[1] Nearly half of Americans report having experienced it.[2] Two-thirds say they have witnessed it.[3] But not everyone is subjected to the same degree of harassment. Certain groups are disproportionately targeted for their identity and profession. Because writers and journalists conduct so much of their work online and in public, they are especially susceptible to online abuse.[4] And among writers and journalists, the most targeted are those who identify as women, BIPOC, LGBTQIA+, and members of religious or ethnic minorities.[5] When voices are silenced and expression is chilled online, public discourse suffers. To ensure that social media becomes safer, more open, and more equitable for all users, platforms like Twitter, Facebook, and Instagram must prioritize curbing online harassment. In this report, PEN America proposes concrete, actionable changes that social media companies should make immediately to the design of their products to better protect people from online abuse—without jeopardizing free expression.

THE DEVASTATING IMPACT OF ONLINE ABUSE

I wasn't prepared emotionally for the abuse I saw on my screen … Now I sometimes avoid reporting on certain topics, or I publish pieces but I just won't post on social media, because I am afraid of the blowback and would rather not deal with it. If I had additional tools to deal with abuse on social media, I would definitely use them.

—Jasmine Bager, journalist

Writers and journalists are caught in an increasingly untenable double bind. They often depend on social media platforms—especially Twitter, Facebook, and Instagram—to conduct, publish and promote their writing. Yet their visibility and the very nature of their work can make them lightning rods for online abuse, especially if they belong to frequently targeted groups and if they cover beats such as feminism, politics, or race.

The consequences of online abuse are dire. It strains mental and physical health,[6] and, in extreme cases, can escalate to physical violence and even murder.[7] Abuse is intended to intimidate and censor. Because the risks to personal health and safety are very real, online harassment has forced some people to avoid writing or speaking about certain subjects, step away from social media,[8] or leave their professions altogether.[9]

When online abuse drives women, LGBTQIA+, BIPOC, and minority writers and journalists to leave industries that are predominantly male, heteronormative, and white, public discourse is impoverished.[10] Individual harms have wider systemic consequences: Online abuse not only undermines equity and inclusion but also inhibits a free press and chills freedom of expression.

SHOUTING INTO THE VOID: INADEQUATE PLATFORM RESPONSE

I report online abuse to the platforms… but it feels like shouting into the void. There's still no transparency or accountability.

—Jaclyn Friedman, writer and founder of Women, Action & the Media

Hate and harassment did not begin with the rise of social media. But because the business model of social media companies is predicated on sustaining user attention and maximizing engagement, their platforms are built to prioritize immediacy, emotional impact, and virality—which amplify abusive behavior.[11] Yet most social media companies have been slow to implement even basic features to fight online abuse—and users have noticed. According to a 2021 study

from Pew Research Center, nearly 80 percent of Americans believe that social media companies are not doing enough to address online harassment.[12]

There is a growing international consensus that the private companies that maintain dominant social media platforms have a responsibility, in accordance with international human rights law and principles, to reduce the harmful impact of abuse while protecting free expression.[13] Calls to regulate social media—from civil society,[14] legislators,[15] and private companies[16]—are mounting. Legislative and regulatory solutions are critically important, but they are also fraught, complex, and hard to get right without further undermining the safety and free speech of individuals already struggling to be heard online. These efforts will take time, but immediate action is urgently needed.

WHAT CAN PLATFORMS DO NOW TO REDUCE THE BURDEN OF ONLINE ABUSE?

You can't have free expression of ideas if people have to worry that they're going to get doxed or they're going to get threatened.

—Mary Anne Franks, president of the Cyber Civil Rights Initiative and professor of law at the University of Miami

If social media companies are serious about addressing online abuse and making their platforms more open and equitable, they must design and build their products with and for their most vulnerable users. With this in mind, PEN America has rooted our recommendations in the experiences and needs of writers and journalists who identify as women, BIPOC, LGBTQIA+, and as members of religious or ethnic minorities. We contend that if technology companies can better protect and support users who are especially vulnerable to online abuse because of their identity and profession, they can better serve *all* users.

As an organization of writers committed to defending freedom of expression, PEN America views online abuse as a threat to the very principles we fight to uphold. When people stop speaking out and writing about certain topics due to fear of reprisal, everyone loses. At the same time, efforts to combat online harassment that rely too heavily on taking down content, especially given the challenges of implicit bias in both human and automated moderation, risk sweeping up legitimate debate and may further marginalize the very individuals and communities such measures are meant to protect.[17]

In this report, PEN America asks: What can social media companies do *now* to ensure that users disproportionately impacted by online abuse receive protection and support? How can social media companies build safer spaces online? How can technology companies, from giants like Facebook and Twitter to small startups, design in-platform features and third-party tools that empower targets of abuse and their allies and disarm abusers? What's working, what can be improved, and where are the gaps?

We propose concrete, actionable changes that technology companies can and should make immediately to the design of their products to combat online abuse while safeguarding free expression. Our recommendations include: **proactive measures** that empower users to reduce risk and minimize exposure; **reactive measures** that facilitate response and alleviate harm; and **accountability measures** to deter abusive behavior. We make the case that technology companies need to better serve users facing both day-to-day abuse and more severe attacks, including threats of violence, doxing, and coordinated mobs.[18]

Throughout this report, in laying out our recommendations, we address the tensions that can arise in countering abuse while protecting free expression, and we propose strategies to mitigate unintended consequences. While the challenges baked into reducing online harms are real, technology companies have the resources and power to implement solutions. Writers, journalists, and other vulnerable users have, for too long, endured relentless abuse on the very social media platforms that they need to do their jobs. It's time for technology companies to step up.

RECOMMENDATIONS

EMPOWERING TARGETED USERS AND THEIR ALLIES

If you're going to be a journalist, there is an expectation to be on social media. [Yet] there are not a lot of resources to protect you. No matter what I say about race, there will be some blowback. Even if I say nothing, when my colleague who is a white man takes positions on racism, trolls come after me on social media.

—Jami Floyd, senior editor, Justice and Race Unit, New York Public Radio

PROACTIVE MEASURES: REDUCING RISK AND EXPOSURE

Proactive measures protect users from online abuse before it happens or lessen its impact by giving its targets greater control to reduce risk and calibrate exposure.

Platforms should:

- Build **shields** that enable users to proactively filter abusive content (across feeds, comments, direct messages, etc.) and quarantine it in a **dashboard**, where they can review and address it with the help of trusted allies.
- Develop robust, intuitive, user-friendly features to fine-tune privacy and security settings, including:
  - Enabling users to customize and save multiple, distinct configurations of settings as **safety modes** that can be activated with one click.
  - Providing users with **visibility snapshots** that show, in real time, how adjusting settings affects reach.
- Make it easier to create and maintain boundaries between personal and professional **online identities**, to migrate or share audiences between those identities, and to easily switch back and forth between them.
- Provide users with robust, integrated features to manage their personal **account histories**, including the ability to search through old posts, review them, make them private, delete them, and archive them—individually and in bulk.
- Enable users to assemble **rapid response teams** of trusted allies and to **delegate account access**, so that those allies can jump in to provide targeted assistance, including mobilizing supportive communities and helping to document, block, mute, and report abuse.

REACTIVE MEASURES: FACILITATING RESPONSE AND ALLEVIATING HARM

Reactive measures, such as blocking and muting to limit interaction with abusive content, can mitigate the harms of online abuse once it is underway.

Platforms should:

- Create an **SOS button** that users can instantly activate to trigger additional in-platform protections and an **emergency hotline** (phone or chat) providing personalized, trauma-informed support in real time.
- Create a **documentation** feature that allows users to record evidence of abuse quickly and easily—capturing screenshots, hyperlinks, and other publicly available data automatically or with one click. Such evidence is critical for communicating with employers, engaging with law enforcement, and pursuing legal action where appropriate.
- Improve and standardize features that help users limit contact with abusive content and accounts, including:
  - **Blocking**, which cuts off contact and communication with abusers;[19]
  - **Muting**, which allows users to hide abusive content from themselves but not from other users;[20] and
  - **Restricting**, which allows users to hide abusive content from all users without alerting the abuser.[21]
- Revamp **reporting** features to ensure they are more user-friendly and trauma-informed. Specifically, platforms should:
  - Create a streamlined, flexible, and responsive report management system, including enabling users to create and edit drafts, add context, and combine multiple reports.
  - Ensure clarity and consistency between reporting features and policies, including providing easy access to rules in real time.

    - Add bulk reporting in recognition of the coordinated nature of harassment campaigns.
    - Create a formal, publicly known appeals or escalation channel for content that is reported as abusive but not taken down.
- Build robust, user-friendly, and easily accessible **anti-abuse help centers** and support the development of promising new third-party tools designed to counter online abuse—especially those built by and for women, BIPOC, and LGBTQIA+ technologists with firsthand experience of harassment.

DISARMING ABUSIVE USERS

Online abuse cannot be addressed solely by creating tools and features that empower the targets of abuse and their allies. Platforms must also actively discourage abuse and hold abusive users accountable. Efforts to deter abuse, however, need to be balanced against competing priorities: They must protect critical speech and prevent the silencing of legitimate dissenting viewpoints, which may include heated debate that does not rise to the level of abuse, as well as humor, satire, and artistic expression.[22] To that end, PEN America's recommendations seek to disarm abusive users without unduly increasing the platforms' power to police critical speech, which threatens all users' free expression rights.

Platforms should:

- Make their **rules**—and the consequences for breaking them—easily accessible **in real time** from directly within the primary user experience, rather than on separate websites. They should deploy all available design elements—including nudges, labels, and contextual clues—to surface relevant rules and encourage policy checkups.
- Create a **transparent system of escalating penalties** for abusive behavior—which should include warnings, strikes, temporary functionality limitations, and suspensions, as well as content takedowns and account bans—and spell out these penalties for users every step of the way.
- Continue to experiment with proactive **nudges** that encourage users to revise abusive content before they post it. Equally important, platforms should study the efficacy of nudges in curbing abuse, publish their findings, and give outside researchers access to the data they need to assess these features independently.
- Revamp the **appeals process** for users whose content or accounts have been taken down, restricted, or suspended. Specifically, platforms should:
    - Communicate clearly and regularly with users at every step.
    - Allow users to add context when they appeal.
    - Ensure that humans review the appealed content.
    - Create a formal, adequately resourced escalation channel for expediting appeals in time-sensitive cases of malicious or inaccurate content or account takedowns.

MITIGATING RISK

No single strategy to fight online abuse will be perfect or future-proof. Any feature intended to combat online abuse is susceptible to gaming and weaponization.[23] In many cases, the difference between an effective strategy and an ineffective or overly restrictive one depends not only on policies but also on the specifics of how features are designed and whom they prioritize and serve. Ensuring that systems are designed to empower users rather than simply prohibit bad behavior can help mitigate those risks, preserving freedom while also becoming more resilient to evolving threats.

PEN America believes that most users are entitled to control who can see and interact with their content, to limit communications with other accounts, especially those engaging in abuse, and to manage their personal account histories. In the case of public officials and entities using social media for official purposes, however, the situation is more complicated. Some features that can mitigate online abuse pose tensions from the standpoint of accountability and transparency. Throughout the full report, we flag risks and propose mitigation strategies.

METHODOLOGY

The recommendations in this report are based on in-depth qualitative research, including over 50 interviews and a comprehensive literature review, and on the extensive experience that PEN America has gleaned through our Online Abuse Defense program, which has reached over 250,000 journalists, writers, editors, academics, lawyers, activists, and other users facing harassment. Specifically, this report:

- Is rooted in the experiences and needs of people disproportionately targeted online because of their identity and/or profession—specifically, writers and journalists whose work requires a public presence online, especially those who identify as women, BIPOC (Black, indigenous, and people of color), LGBTQIA+ (lesbian, gay, bisexual, transgender, queer, intersex, and asexual), and/or as members of religious or ethnic minorities.[24]
- Focuses on Twitter, Facebook, and Instagram because they are the platforms on which United States–based writers and journalists rely most heavily for their work,[25] and the platforms on which United States–based users report experiencing the most abuse.[26]
- Focuses on the product design of social media platforms, specifically in-platform features and third-party tools that address online abuse.
- Addresses the United States context, where PEN America's expertise in online abuse is strongest. We recognize, however, that online abuse is a global problem and endeavor to note the risks and ramifications of applying strategies conceived in and for the United States internationally.[27]
- Recognizes that online abuse can be multidirectional and that the lines between abuser, target, and ally are not always clear-cut. Because a user can be either an abuser or a target at any time, features designed to address online abuse must approach it as a behavior—not an identity.

ENDNOTES

1. PEN America defines online abuse as the "severe or pervasive targeting of an individual or group online with harmful behavior." "Defining 'Online Abuse': A Glossary of Terms," Online Harassment Field Manual, accessed January 2021

2. "Online Hate and Harassment Report: The American Experience 2020," ADL, June 2020; see also Emily A. Vogels, "The State of Online Harassment," Pew Research Center, January 13, 2021

3. Maeve Duggan, "Online Harassment 2017: Witnessing Online Harassment," Pew Research Center, July 11, 2017

4. Michelle P. Ferrier, "Attacks and Harassment: The Impact on Female Journalists and Their Reporting (Rep.)," IWMF/TrollBusters, 2018; "Why journalists use social media," NewsLab, 2018

5. For impact on female journalists internationally, see Ibid. and Julie Posetti et al., "Online violence Against Women Journalists: A Global Snapshot of Incidence and Impacts," UNESCO, December 1 2020; For impact on women and gender nonconforming journalists in the U.S. and Canada, see: Lucy Westcott, "'The threats follow us home': Survey details risks for female journalists in U.S., Canada," CPJ, September 4, 2019; For impact on women of color, including journalists, see: "Troll Patrol Findings," Amnesty International, 2018

6. Michelle P. Ferrier, "Attacks and Harassment: The Impact on Female Journalists and Their Reporting," IWMF/TrollBusters, 2018; Lucy Westcott, "'The Threats Follow Us Home': Survey Details Risks for Female Journalists in U.S., Canada," Committee to Protect Journalists, September 4, 2019

7. According to a recent global study of female journalists conducted by UNESCO and the International Center for Journalists (ICFJ), 20 percent of respondents reported that the attacks they experienced in the physical world were directly connected with online abuse. Julie Posetti et al., "Online violence Against Women Journalists: A Global Snapshot of Incidence and Impacts," UNESCO, December 1 2020; "The Committee to Protect Journalists has reported that 40 percent of journalists who are murdered receive threats, including online, before they are killed." Elisabeth Witchel, "Getting Away with Murder"

8. "Online Harassment Survey: Key Findings," PEN America, 2017; Mark Lieberman, "A growing group of journalists has cut back on Twitter, or abandoned it entirely," Poynter Institute, October 9, 2020

9. Michelle P. Ferrier, "Attacks and Harassment: The Impact on Female Journalists and Their Reporting," IWMF/TrollBusters, 2018

10. "What Online Harassment Tells Us About Our Newsrooms: From Individuals to Institutions," Women's Media Center, 2020

11. Amit Goldenberg and James J. Gross, "Digital Emotion Contagion," Harvard Business School, 2020; Luke Munn, "Angry by design: toxic communication and technical architectures," Humanities and Social Sciences Communications 7, no. 53 (2020); Molly Crockett, "How Social Media Amplifies Moral Outrage," The Eudemonic Project, February 9 2020

12. Emily A. Vogel, "The State of Online Harassment," Pew Research Center, January 13, 2021

13. Susan Benesch, "But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies," Yale Journal on Regulation Online Bulletin 3 (September 14, 2020)

14. "Toxic Twitter—A Toxic Place for Women," Amnesty International, 2018; Eva Galperin and Dia Kayyali, "Abuse and Harassment: What Could Twitter Do?," Electronic Frontier Foundation, February 20, 2015

15. Davey Alba, "Facebook Must Better Police Online Hate, State Attorneys General Say," The New York Times, August, 5, 2020

16. Kelly Tyko, "Facebook advertising boycott list: Companies halting ads include Unilever, Coca-Cola, Verizon, Ben & Jerry's," USA Today, June 27, 2020

17. Mallory Locklear, "Facebook is still terrible at managing hate speech," Engadget, August 3, 2017; Tacey Jan, Elizabeth Dwoskin, "A White Man Called Her Kids the N-Word. Facebook Stopped Her from Sharing it." The Washington Post, July 31st, 2017

18. PEN America defines doxing as the "publishing of sensitive personal information online—including home address, email, phone number, social security number, photos, etc.—to harass, intimidate, extort, stalk, or steal the identity of a target." "Defining 'Online Abuse': A Glossary of Terms," Online Harassment Field Manual, accessed January 2021

19. Kat Lo, "Toolkit for Civil Society and Moderation Inventor," Meedan, November 18, 2020

20. Ibid

21. Katy Steinmetz, "What to Know About Restrict, Instagram's New Anti-Bullying Feature," Time, July 8, 2019

22. Sam Biddle, "Facebook Lets Vietnam's Cyberarmy Target Dissidents, Rejecting a Celebrity's Plea," The Intercept, December 12, 2020; Russell Brandom, "Facebook's Report Abuse button has become a tool of global oppression," The Verge, September 2, 2014; Ariana Tobin, Madeline Varner, Julia Angwin, "Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up," ProPublica, December 18, 2017; Katie Notopoulos, "How Trolls Locked My Twitter Account For 10 Days, And Welp," BuzzFeed News, December 2, 2017; Tracey Jan, Elizabeth Dwoskin, "A White Man Called Her Kids the N-Word. Facebook Stopped Her from Sharing it." The Washington Post, July 31st, 2017

23. Katie Notopoulos, "How Trolls Locked My Twitter Account For 10 Days, And Welp," BuzzFeed News, December 2, 2017; Tracey Jan, Elizabeth Dwoskin, "A White Man Called Her Kids the N-Word. Facebook Stopped Her from Sharing it." The Washington Post, July 31st, 2017; Russell Brandom, "Facebook's Report Abuse button has become a tool of global oppression," The Verge, September 2, 2014; Sam Biddle, "Facebook Lets Vietnam's Cyberarmy Target Dissidents, Rejecting a Celebrity's Plea," The Intercept, December 12, 2020; Ariana Tobin, Madeline Varner, Julia Angwin, "Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up," ProPublica, December 18, 2017

24. "Online Hate and Harassment Report: The American Experience 2020," ADL, June 2020

25. "Michelle P. Ferrier, "Attacks and Harassment: The Impact on Female Journalists and Their Reporting (Rep.)," IWMF/TrollBusters, 2018; "Why journalists use social media," NewsLab, 2018; "2017 Global Social Journalism Study," Cision, accessed February 19, 2021

26. "Online Hate and Harassment Report: The American Experience 2020," ADL, June 2020; see also Emily A. Vogels, "The State of Online Harassment," Pew Research Center, January 13, 2021

27. "Activists and tech companies met to talk about online violence against women: here are the takeaways," Web Foundation, August 10, 2020