MORE HATE, FEWER PROTECTIONS.

Harmful Content on Meta's Platforms in the Wake of Rollbacks, According to Users

ULTRAVIOLET glaad ALL OUT

TABLE OF CONTENTS

EXECUTIVE SUMMARY 3

INTRODUCTION 7

METHODOLOGY 11

FINDINGS 13

DISCUSSION 23

ACKNOWLEDGEMENTS

Deep gratitude to Nicole Regalado, Jenni Olson, and Justin Lessner for your crucial support of this project. Thank you to the All Out local campaigners who translated this survey into 5 languages, making our survey more accessible across the globe. And most importantly, thank you to every user who took the time to complete this survey. Your perspectives are invaluable.

AUTHORS

Jenna Sherman,
Campaign Director, UltraViolet

Ana Clara Toledo,
Senior Campaigns Manager in Latin America, All Out

Leanna Garfield,
Senior Manager, Social Media Safety Program, GLAAD

Contact: media@weareultraviolet.org

EXECUTIVE SUMMARY

On January 7, 2025, Meta announced sweeping changes to its content moderation policies, including the end of third-party fact-checking in the U.S., and rollbacks to its hate speech policy globally that removed protections for women, people of color, trans people, and more.[1,2]

These policy shifts signified a dramatic reversal of content moderation standards the company had built over nearly a decade.

Experts immediately warned about the risks that these rollbacks posed to the over 3 billion people globally who are active on Facebook, Instagram, or Threads, particularly historically marginalized users.[3] In the absence of data from Meta on the impacts of these policy shifts on users, we decided to go straight to users—arguably the experts on what content shows up in feeds on Meta platforms—to assess if and how harmful content is manifesting on Meta platforms, and to better understand the real-world effects of Meta's recent policy changes, specifically among vulnerable users including women, LGBTQ people, and people of color.

Among our survey population of approximately 7,000 active users, we found stark evidence of increased harmful content, decreased freedom of expression, and increased (self-)censorship.

In essence, we found evidence of Meta users—especially users who belong to what Meta defines as a protected characteristic group—experiencing a reality on Meta platforms that runs counter to the promise Meta CEO Mark Zuckerberg made when he announced the sweeping changes.[4,5]



Since the January 2025 policy rollbacks:

1 in 6

of respondents reported being the **victim of some form of gender-based or sexual violence** on Meta platforms.

92%

of respondents say they are **concerned about harmful content increasing** on Meta platforms.

72%

of respondents report that **harmful content targeting protected groups** has increased.

92%

of respondents say they feel **less protected from being exposed to or targeted** by harmful content on Meta platforms.

OVER 1/4

of respondents report being **targeted directly with harmful content** on a Meta platform.

66%

of respondents have **witnessed harmful content** on Meta platforms.

77%

of respondents describe **feeling less safe** expressing themselves freely.



We urge Meta to hire an independent third-party to formally analyze changes in harmful content facilitated by the policy changes—centering the perspective of users—and

URGENTLY REINSTATE PROTECTIONS AND WIDESPREAD CONTENT MODERATION FOR USERS.

Users deserve online spaces where they can feel safe and thrive. They cannot do that when continually targeted with hate and harassment on the basis of who they are.

INTRODUCTION

INTRODUCTION

Most of us go on social media platforms looking for connection and community. A big part of what allows us to do so in relative safety is the fact that

a massive amount of harmful content--from violent depictions to hate speech to disturbing images--has already been taken down by human or machine reviewers because it is in violation of platform policies.

Since the early 2000s, social media platforms have been experimenting with how to best protect users from potentially harmful content online, including hate speech and disinformation. The balance has never been perfect, but most platforms—including Meta (formerly named Facebook, and the owner of Facebook, Instagram, and Threads)—have generally worked toward improving their trust and safety protocols. Meta's content moderation system, deployed in 2016, involved tens of thousands of reviewers from around 119 countries working to keep people like us safe on Facebook and Instagram.[6] Though imperfect, this system both helped keep users safe and set a standard of care for mitigating harmful content online, which became increasingly crucial as our lives grew more intertwined with the digital world.

This year, Meta has made pivotal shifts to their content moderation practices,[7] what some have called a "MAGA makeover."[8] On January 7, 2025, Meta announced sweeping changes to its existing policies, including the end of third-party fact-checking in the U.S. and rollbacks to its hate speech policy globally.[9] In a blog post titled "More Speech and Fewer Mistakes," Joel Kaplan, Meta's chief global affairs officer, detailed several key changes:

The dissolution of its fact-checking program in the U.S.; the removal of policies on "immigration, gender identity, and gender"; and the cessation of "proactive" enforcement of some policies on harmful content.[10]

While the changes to enforcement and fact-checking would "be expanded beyond the U.S." at a future date, Kaplan said, those to the hate speech policy had "been implemented worldwide immediately."[11] (According to a recent analysis conducted by the Center for Countering Digital Hate, these rollbacks were predicted to affect 97% of Meta's enforcement, which, according to its own data, would result in nearly 277 million pieces of hate speech and harmful content flooding Meta platforms unchecked each year.)[12]

The wide-ranging changes include more allowances for hate and harassment targeting historically marginalized groups, including women, people of color, LGBTQ people, and more,[13] who already face disproportionate levels of abuse online.[14] In addition to the removal of protections from its newly renamed Hateful Conduct policy (previously the "Hate Speech Community Standard"), Meta also added terms ("transgenderism" and "homosexuality") that are well-known to convey animus against LGBTQ people, especially transgender and gender-nonconforming individuals.[15]

Shortly after the announcement, two news outlets leaked Meta's revised internal content moderation guidelines, which revealed that the company would now allow other dehumanizing statements, such as:

"Immigrants are grubby, filthy pieces of shit;"[16] "Black people are more violent than Whites;" "Jews are flat out greedier than Christians;" and "A trans person isn't a he or she, it's an it;" as well as expressly adding the allowance of "allegations of mental illness or abnormality when based on gender or sexual orientation."[17]

These shifts are in direct conflict with Meta's Community Standards, which still state that the company doesn't "allow hateful conduct," defined as "direct attacks against people ... on the basis of what we call protected characteristics (PCs): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease."[18]

These changes mark a significant deviation from Meta's posture towards content moderation over the last near-decade; Zuckerberg initially implemented the fact-checking system in 2016 after Facebook faced criticism for facilitating the spread of disinformation around Donald Trump's first election win. In 2020, following a damning report from auditors and a widespread #StopHateForProfit advertising boycott, the company announced it would begin putting warning labels on violent posts, including by Donald Trump, and removing all pages supporting QAnon.[19,20]

In essence, Zuckerberg and his team were somewhat receptive to public and civil society pressures, and touted the fight against hate speech as a priority, at least in word.

While Meta's content moderation systems were never perfect, and Zuckerberg's stance was always capricious, it's clear we are now in a vastly new era of Meta to the shock and concern of many. The policy rollbacks and subsequent revelations have drawn widespread concern from human rights organizations[21] as well as from Meta's own Oversight Board, which has recommended that Meta "identify how the policy and enforcement updates may adversely impact the rights of LGBTQIA+ people, including minors, especially where these populations are at heightened risk."[22] Civil society groups were also reportedly not consulted on the changes,[23] a significant departure from established protocols that seek to uphold human rights and free expression while reducing the risk of physical violence, discrimination, and other offline consequences.[24] As many have noted,[25] this dismantling of protections—introduced alongside Donald Trump's return to the presidency, mass layoffs,[26] and cuts to company-wide DEI programming[27]—signals a distinct shift in Meta's content moderation strategy, increasing the potential for harm to billions of users around the world.

Meta offers some transparency around its actions on harmful content through quarterly reports, which disclose estimated violation numbers based on the company's own assessments. In its most recent quarterly report, published May 29, 2025, the company reported a rise in violent content alongside a sharp decrease in false flags since the change in content moderation policies.[28,29]

However, these reports are not independently conducted and do not reflect *users'* experiences of targeted hate and harassment, nor the degree to which the company is adequately enforcing its policies.

The data are also not provided, making their analysis a black box. The goal of this public survey was to hear from users about how harmful content is manifesting on Facebook, Instagram, and Threads, and to better understand the real-world effects of Meta's recent shifts.

METHODOLOGY

METHODOLOGY

This survey is a cross-sectional, mixed methods survey targeted at active users of Instagram, Facebook, and Threads.

Specifically, our target audience was what Meta refers to as "protected characteristic groups," which include people targeted based on their race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.[30]

Given that our target audience was protected groups as Meta defines them, we recruited survey participants by email and social media outreach to the audiences of cosponsor organizations. We chose not to distribute the survey more widely or via paid advertising out of concern about skewed results or survey trolling.[31] Due to the specificity of the groups surveyed and the nonrandom nature of the sampling, these results are not generalizable to all Meta users.

In terms of survey scope, we defined "harmful content" as Meta defines it:[32]

Content that involves direct attacks against people based on a protected characteristic.

However, the survey also left room for respondents to further explain their experiences of psychological or physical harm. There were nine required questions and five optional questions.

There was one English-language survey conducted on Jotform, which garnered 5,278 respondents. The English survey was also translated into Portuguese, Spanish, German, Italian, and French, and shared organically through All Out's social channels and mailing lists globally via a Typeform survey, which garnered a total of 1,754 respondents. We translated qualitative testimonies from those languages into English for the purposes of this survey report. Finally, we cleaned the data for any repeat submissions before analyzing.



FINDINGS

FINDINGS

TOTAL COUNTRIES REPRESENTED

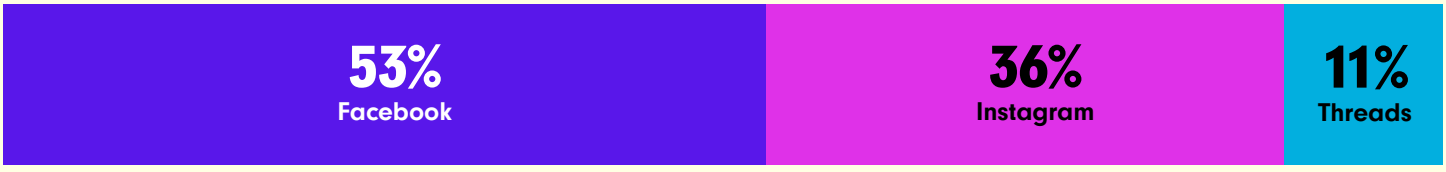
86 NATIONS

AVERAGE AGE

50.5 YEARS

COUNTRIES WITH OVER 10 RESPONDENTS

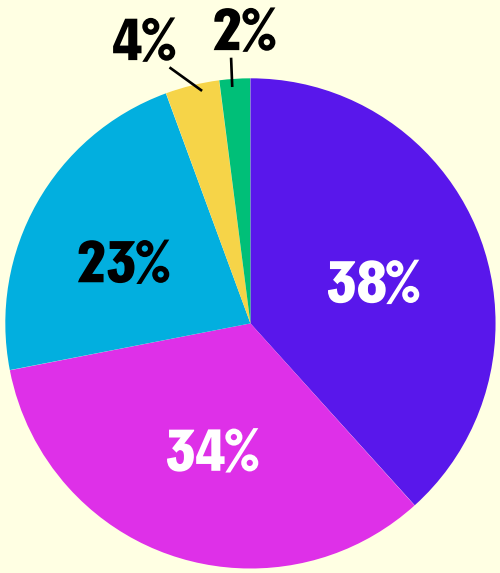
BREAKDOWN OF RESPONDENTS' META PLATFORM USAGE



7,032
Total Respondents

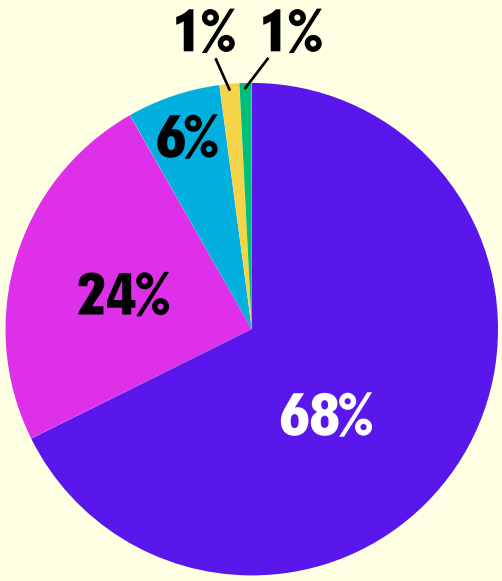
PERCEIVED CHANGES IN HARMFUL CONTENT

From your experience, has harmful content targeting protected groups increased, decreased, or stayed the same since January 2025? Harmful content includes hate speech, like slurs and violent content, and harassment, like bullying and calls for discrimination.



CONCERN ABOUT HARMFUL CONTENT IN THE WAKE OF POLICY ROLLBACKS

Following Meta's recent rollbacks on hate speech and content moderation, how concerned are you about harmful content increasing on Meta platforms?

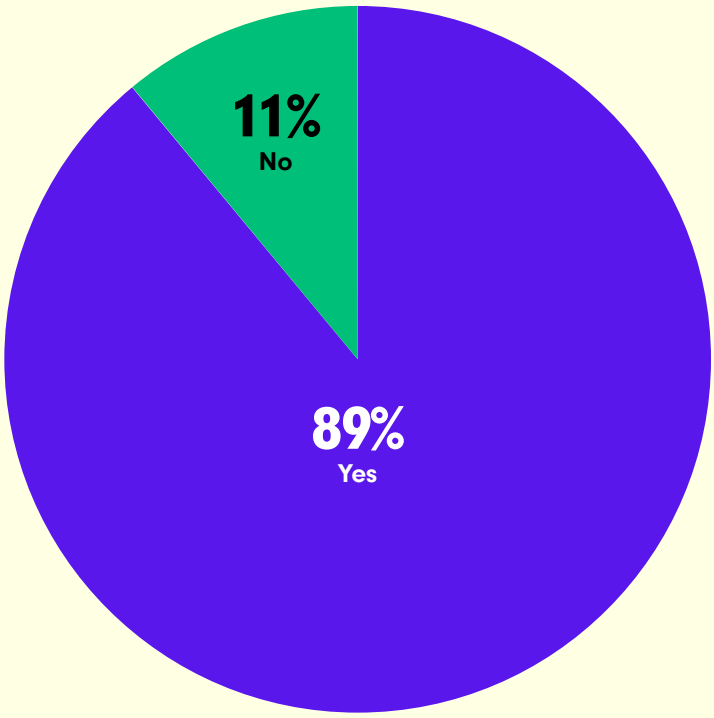


DIRECT EXPERIENCE WITH HARMFUL CONTENT

Have you been the target of any form of harmful content on any Meta platform since January 2025?

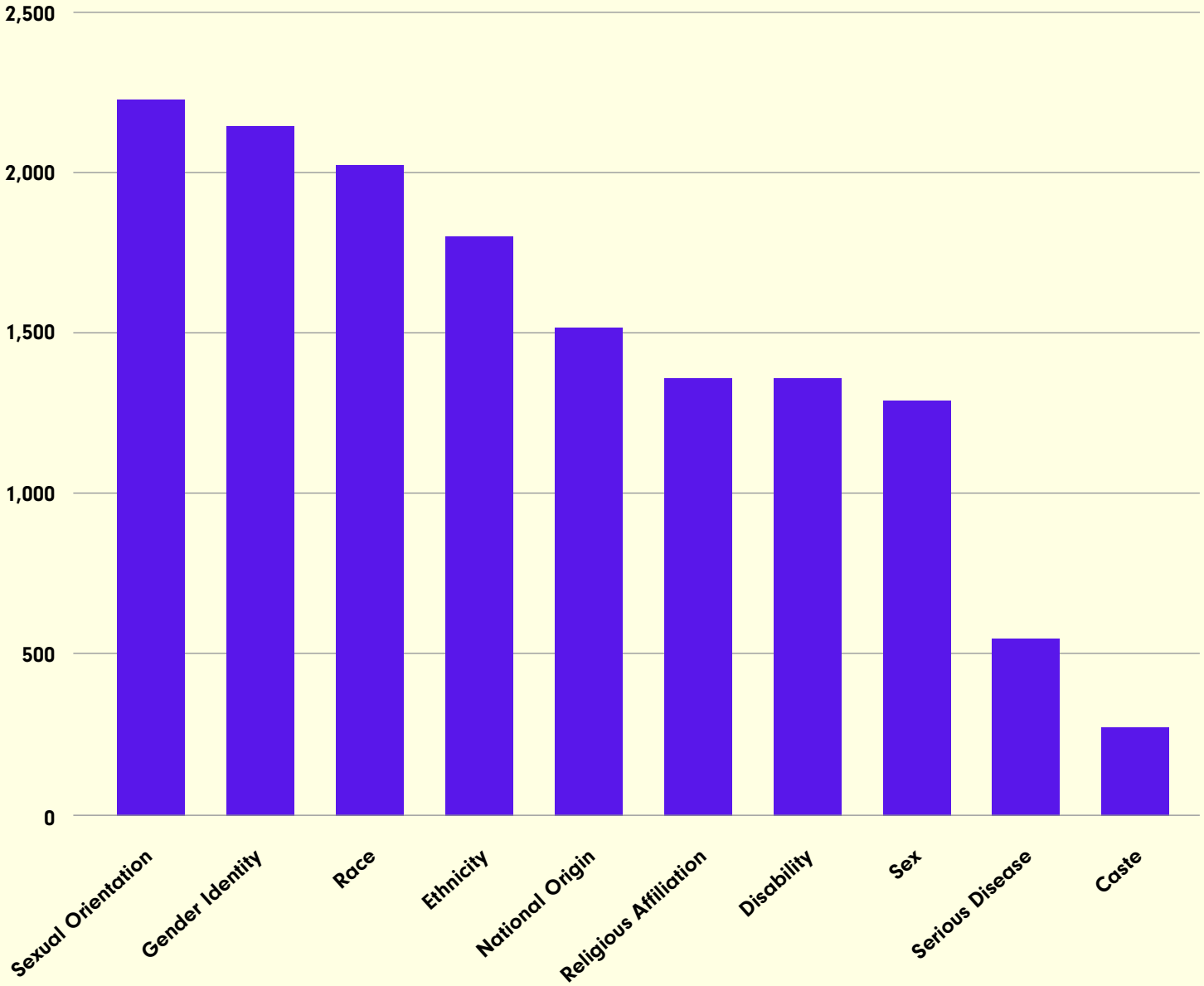
Have you witnessed any form of harmful content on any Meta platform since January 2025?

If yes to either of the above, were any of the attacks targeting you or others due to protected characteristics?



HARMFUL CONTENT BASED ON PROTECTED CHARACTERISTICS (CONT.)

If yes to either of the above, were any of the attacks targeting you or others due to the following protected characteristics? Select all that apply.



The vast majority of these groups reported that harmful content targeting protected groups has increased since the January 2025 policy rollbacks.

75%

of LGBTQ respondents

76%

of women respondents

78%

of respondents who identified as a person of color

WHAT RESPONDENTS ARE SAYING

"Vulgar and sexual remarks towards women. Harassment of people [based on] their races. Members bragging about attacking others who are non-white and how they'll video what they do to them and being cheered on by other members.

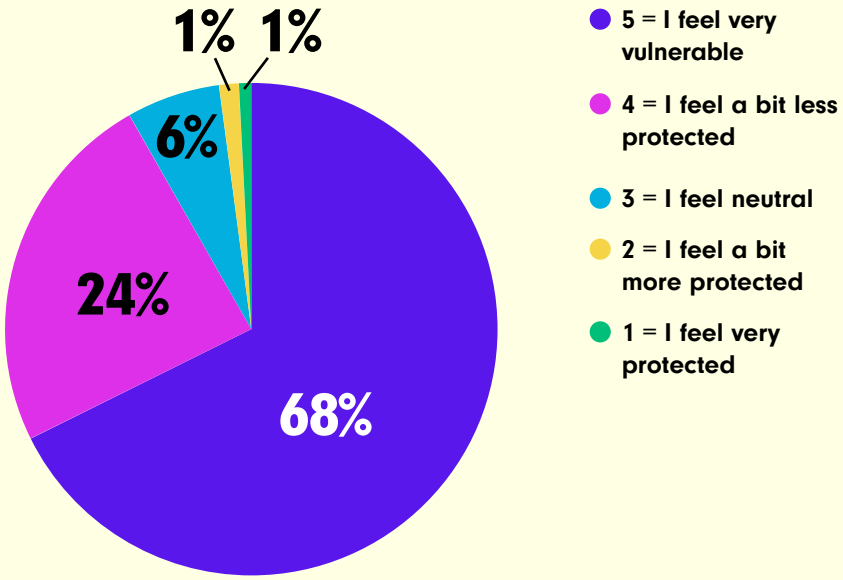
"I received a comment wishing I would be gang raped by Black migrants to 'bring me back to reality,' along with threats of being attacked by Muslim extremists because 'Muslims kill LGBT people.' Reporting these to Facebook was useless – all reports were rejected.

"One night I reported at least 10 comments directly inciting violence towards the LGBT community. FB responded within less than a minute saying that the comments were investigated and they didn't see anythingwrong, and kept the comments up.

"One popular 'meme' is that of someone's feet dangling, wearing socks in the colours of the trans flag with the pronouns 'was/were'. I think it's pretty clear what this is suggesting yet every time I report it I get told it does not breach Meta's rules.

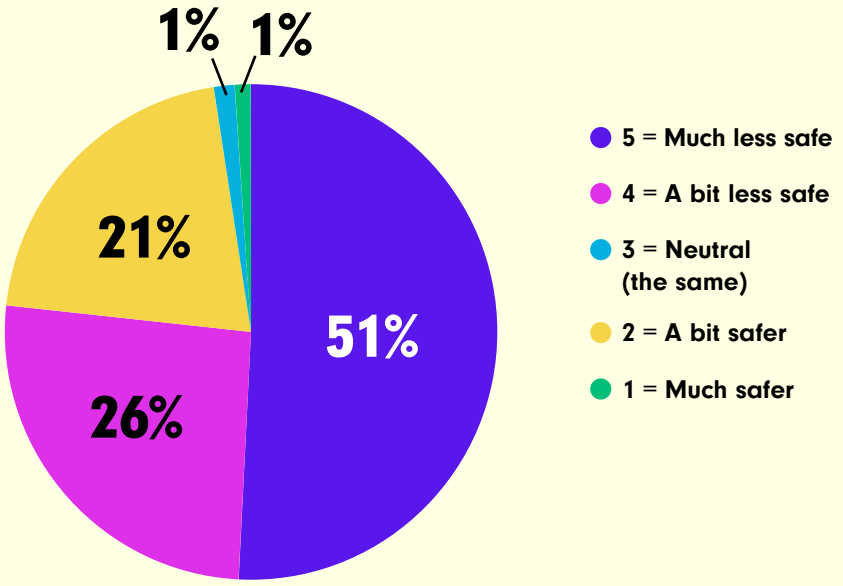
USER PROTECTION IN THE WAKE OF POLICY ROLLBACKS

How well do you feel Meta's new policy changes protect you and all users from being exposed to or targeted by harmful content?



FREEDOM OF EXPRESSION

How safe do you feel to express yourself on Meta's platforms following January 2025?



Protected groups feel less safe expressiong themselves on Meta platforms since the January 2025 policy rollbacks.

81%

of LGBTQ respondents

83%

of women respondents

77%

of respondents who identified as a person of color

WHAT RESPONDENTS ARE SAYING

"I'm scared of commenting against our president and his cronies. I fear I wouldn't be safe.

"Opinions and free speech are being targeted in very hateful ways: demeaning; education shaming; gender shaming – you name it.

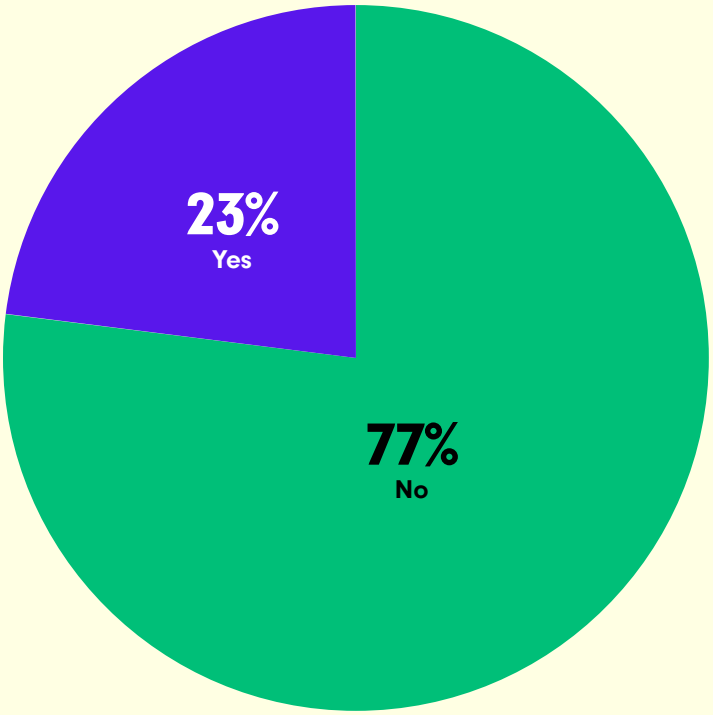
"After January, I received two notifications from Instagram questioning whether it was really me posting pro-Palestine, feminist, and queer-supportive content. I was threatened with temporary or permanent suspension, something that never happened before.

"I have had my posts 'fact checked' because I am a female who is not in support of Trump. Hate speech from Trump supporters is tolerated and likely encouraged and supported.

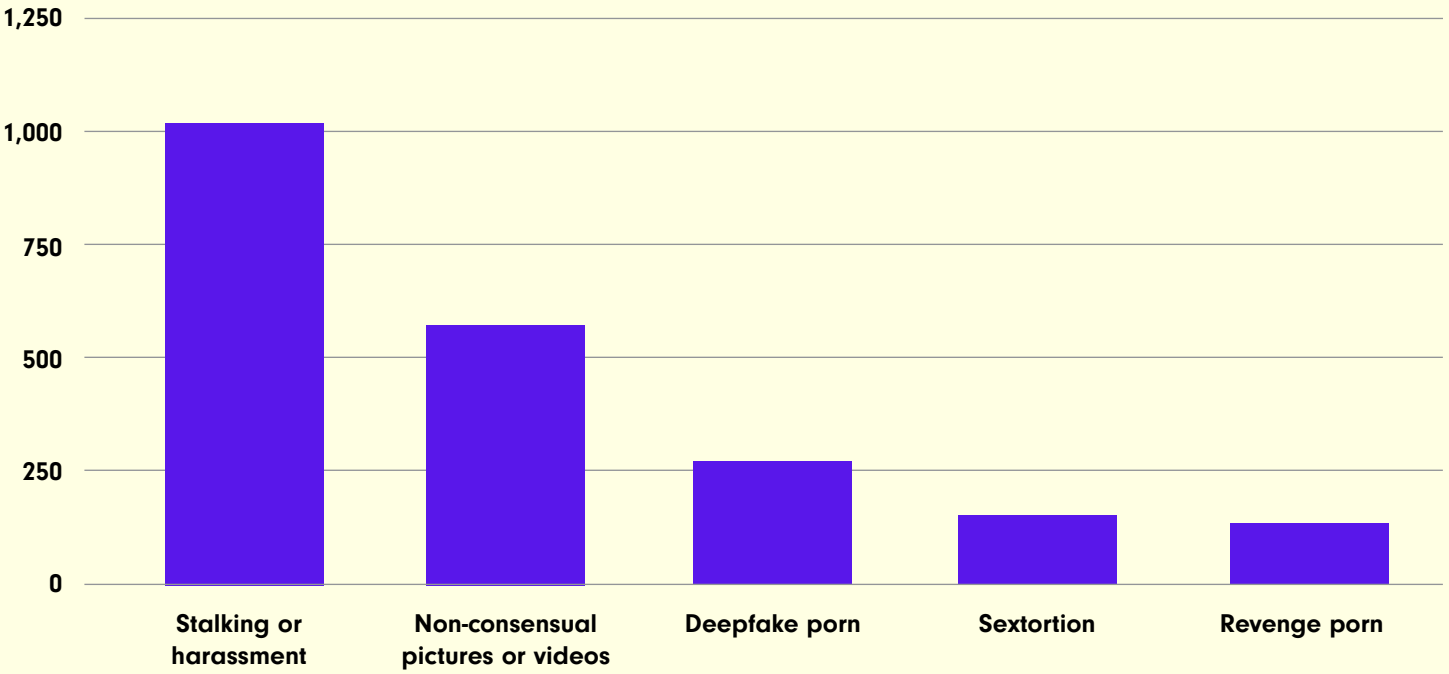
"I have come across TONS of death and sexual assault threats. I have seen far too many heartbreaking videos where these women are scared for their lives.

"I've been harassed on these platforms by men for simply expressing an opinion. I've also been stalked by men on the platform.

Since January 2025, have you been a victim of any form of gender-based or sexual-violence on Meta platforms?



Since January 2025, have you been the victim of any of following forms of gender-based or sexual violence on Meta platforms? Select all that apply.



Protected groups reported significantly higher rates of gender-based or sexual violence.

27%

of LGBTQ respondents

35%

of respondents who identified as a person of color

WHAT RESPONDENTS ARE SAYING

"I have been doxxed, stalked, and threatened with physical harm from people on FB."

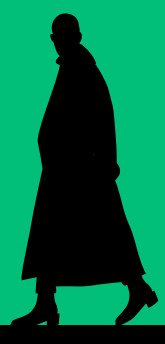
"I've been told women should know their place if we want to support America. I've been sent DMs requesting contact based on my appearance. I've been primarily stalked due to my political orientation."

"Allowing rape threats and threats of domestic violence makes everyone who isn't a man feel threatened and unsafe."

"I was told that as a woman I should be 'properly fucked by a real man' to 'fix my head' regarding gender equality and LGBT+ rights."

"Digital stalking based on my sexual orientation and gender identity was a deeply invasive act that weaponizes technology to threaten, harass, and silence me, transforming their online existence into a battleground of fear and vulnerability, undermining my safety and well-being."

DISCUSSION



DISCUSSION: IMPLICATIONS

This survey is the first of its kind to assess user-reported experiences with harmful content and perceived changes in harmful content in the wake of the sweeping policy rollbacks implemented by Meta in January 2025. Those rollbacks included the end of third-party fact-checking in the U.S. and rollbacks to its hate speech policy globally. The results clearly demonstrate a sharp perceived increase in harmful content, high concern among users about increasing harmful content, a decreased sense of freedom to safely self-express, a decreased feeling among users that they are protected on Meta platforms, and high self-reported rates of users being victims of hateful content and gendered violence.

Notably, findings from this survey demonstrate that the majority of the harmful content respondents witnessed or experienced on Meta platforms was targeted on the basis of what Meta itself refers to as protected characteristics: "race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease." We found the protected characteristics most targeted were sexual orientation and gender identity, closely followed by race, ethnicity, and national origin. These quantitative findings mirror our qualitative findings, in which a large number of users reported witnessing or experiencing an increase in online attacks against women, people of color, and LGBTQ people.

Overall, these results should set off alarm bells for Meta, and for all social media companies, as evidence of what can happen when content moderation systems—particularly long-standing ones—are dismantled and when hate speech and harmful content policy guidance is weakened under the guise of free expression.

As evidenced by the findings, vulnerable users, in fact, feel less safe expressing themselves on Meta platforms under the updated policies that Meta frames as a return to free speech on its platforms.

When Meta announced its widespread policy changes in January 2025, human rights experts were immediately concerned, particularly given the one-two punch of a gutted content moderation system with a weakened hateful speech policy–a policy that effectively sets the norms for the platform.

In essence, Meta undermined not just the enforcement mechanisms (third-party fact-checking) but the rules (the hateful conduct policy), which was certain to result in changes across its platforms for over 3 billion global users.

We are still working to piece together the extent of those changes and how they translate to offline harms. This survey is a first step, and it's clear the impacts are significant.

We also know that while Meta's policy shifts on harmful content will have global effects, they will not be felt in the same way by everyone. In countries and communities where marginalized groups already face higher vulnerability, the impact will be even more severe. This is especially true for communities and entire countries where LGBTQ individuals are criminalized or where women and queer people receive minimal protections.

That's because extant research is clear that online harms can easily translate to offline violence.[33,34]

In Latin America, for instance, where rates of violence against LGBTQ individuals are among the highest in the world, the effects of unchecked hate are already visible. In our survey, Colombian respondents shared stories of the attacks they had endured and the hatred they had witnessed against trans and other LGBTQ people following the recent murder of Sara Millerey González, a trans woman whose brutal killing was filmed and spread across social media. As our lives become increasingly intertwined with the digital world, it's more imperative than ever to take online violence seriously. Instead, Meta is effectively relinquishing its responsibility to mitigate hateful content online, putting our lives at risk in the process.

Findings from this survey already clearly indicate those real risks and the disparate ways in which they are felt. For some survey respondents, Meta's policy shifts are already resulting in a more hateful, spammy, violent online experience. For others, they have led to hate that threatens their very existence.

As Meta is the largest social media company, with billions of users worldwide, the implications of its policy rollbacks for humanity cannot be overstated.

We urge Meta to formally analyze changes in harmful content caused by the policy changes, and to urgently reinstate protections and widespread content moderation for users.

Users deserve online spaces where they can feel safe and thrive. They cannot do that when continually targeted with hate and harassment on the basis of who they are.

DISCUSSION: LIMITATIONS

There are a number of limitations to this survey that demonstrate the need for further investigation and research.

First, as mentioned in the methodology section, this survey is not a scientific research study and was conducted with nonrandom sampling. While the nonrandom survey outreach was intentional, to reach people who belong to what Meta refers to as a "protected characteristic group," we cannot extrapolate the findings to all Meta users.

Next, though our survey respondents were geographically diverse, representing 86 different countries, the majority of respondents were from the United States, the United Kingdom, or Canada, making the findings less representative of the global majority.

This limitation is particularly notable given the large number of Meta users based in the Global South, and because of the different and often disproportionate risks that Global South users face online and offline.[35,36]

Third, the anonymous nature of the survey meant that we could not wholly account for duplicates or corroborate stories. And finally, self-reported data is always subject to the limitations of bias and the subjectivity of interpretation; though we provided a definition of "harmful content" (as defined by Meta), different individuals may interpret that term differently. At the same time, users' self-reported data–separated from company influence–reflects information that empirical, company-provided data cannot: specifically, how users actually perceive harmful content to be changing on Meta platforms, and the impact they experience from that harmful content on their daily lives.

ENDNOTES

[1] Meta, "More Speech and Fewer Mistakes," January 7, 2025, https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

[2] The New York Times, "Meta to End Fact-Checking Program in Shift Ahead of Trump Term," January 7, 2025, https://www.nytimes.com/2025/01/07/technology/meta-fact-checking-facebook.html

[3] Meta, "About Meta," accessed May 21, 2025, https://www.meta.com/about/company-info/

[4] Meta, "Hateful Conduct," accessed May 21, 2025, https://transparency.meta.com/policies/community-standards/hateful-conduct/

[5] Meta, "More Speech and Fewer Mistakes," January 7, 2025, https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

[6] Poynter, Meta is ending its third-party fact-checking partnership with US partners. Here's how that program works., January 7, 2025, https://www.poynter.org/fact-checking/2025/meta-ends-fact-checking-community-notes-facebook/

[7] CNN, "Mark Zuckerberg's MAGA makeover will reshape the entire internet," January 7, 2025, https://www.cnn.com/2025/01/07/media/mark-zuckerberg-meta-fact-checking-analysis/index.html

[8] The New York Times, "What's Behind Meta's MAGA Makeover?" January 8, 2025, https://www.nytimes.com/2025/01/08/technology/meta-facebook-trump-mark-zuckerberg.html

[9] Meta, "More Speech and Fewer Mistakes," January 7, 2025, https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

[10] Ibid.

[11] Kaplan: "Some of the changes, like the changes to our Community Standards on Hateful Conduct, those have been implemented worldwide immediately. The enforcement changes are rolling out in the U.S. first, and those, too, we'll iterate and make sure that we've got it right and are doing it responsibly, and then those will be expanded beyond the U.S. as well. But the Community Notes system is the one I was focused on in saying that it's probably going to take us this year to get it right in the U.S., and then we'll expand from there." The quote appears at 06:12 in a video posted on Meta on February 7, 2025, under the headline "Joel Kaplan on EU Regulation and Innovation",(https://about.fb.com/news/2025/02/joel-kaplan-on-eu-regulation-and-innovation/)

[12] Center for Countering Digital Hate, "Meta Policy Changes Threaten 97% of its Hate Speech Enforcement," February 24, 2025, https://counterhate.com/blog/meta-policy-changes-press-release/

[13] Wired, "Meta Now Lets Users Say Gay and Trans People Have 'Mental Illness,'" January 7, 2025, https://www.wired.com/story/meta-immigration-gender-policies-change/

[14] Pew Research Center, "The State of Online Harassment," January 13, 2021, https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

[15] GLAAD, "Online Anti-LGBTQ Hate Terms Defined: 'Transgenderism,'" https://glaad.org/transgenderism-definition-meaning-anti-lgbt-online-hate/

[16] The Intercept, "Leaked Meta Rules: Users are free to post 'Mexican immigrants are trash!' or 'Trans people are immoral,'" January 9, 2025, https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/

[17] Platformer, "Inside Meta's dehumanizing new speech policies for trans people", January 9, 2025, https://www.platformer.news/meta-new-trans-guidelines-hate-speech/

[18] Meta, "Hateful Conduct" policy, January 7, 2025, https://transparency.meta.com/policies/community-standards/hateful-conduct/

[19] Vanity Fair, "Facebook Puts a Label on Trump's Lies, Calls It A Day," November 17, 2020, https://www.vanityfair.com/news/2020/11/facebook-misinformation-labels-not-working-trump

[20] The Guardian, "Facebook to ban QAnon-themed groups, pages and accounts in crackdown," October 6, 2020, https://www.theguardian.com/technology/2020/oct/06/qanon-facebook-ban-conspiracy-theory-groups

[21] Amnesty International, "Meta's new content policies risk fueling more mass violence and genocide," February 17, 2025, https://www.amnesty.org/en/latest/news/2025/02/meta-new-policy-changes/

[22] Oversight Board, "Gender Identity Debate Videos," April 23, 2025, https://www.oversightboard.com/decision/bun-1ynnk264/

[23] Center for Democracy and Technology, "Letter from Meta Civil Rights Advisory Group Members on Grave Concerns with Content Policy Changes," January 14, 2025, https://cdt.org/insights/letter-meta-civil-rights-advisory-group-members-grave-concerns-content-policy-changes/

[24] UNESCO, "Guidelines for the Governance of Digital Platforms," 2023, https://unesdoc.unesco.org/ark:/48223/pf0000387339

[25] Tech Policy Press, "Meta's Content Moderation Changes are Going to Have a Real World Impact. It's Not Going to be Good," January 9, 2025, https://www.techpolicy.press/metas-content-moderation-changes-are-going-to-have-a-real-world-impact-its-not-going-to-be-good/.

[26] Newsweek, "Meta Unleashes Mass Layoffs," February 10, 2025, https://www.poynter.org/fact-checking/2025/meta-ends-fact-checking-community-notes-facebook/

[27] Axios, "Exclusive: Meta kills DEI programs," January 10, 2025, https://www.axios.com/2025/01/10/meta-dei-programs-employees-trump

[28] Meta, "Integrity Reports, First Quarter 2025," May 29, 2025, https://transparency.meta.com/integrity-reports-q1-2025

[29] Business Insider, "Meta says online harassment is up and false flags are down following a change in content moderation policies," May 29, 2025, https://www.businessinsider.com/meta-says-online-harassment-is-up-after-content-moderation-changes-2025-5

[30] Meta, "Hateful Conduct," accessed May 21, 2025, https://transparency.meta.com/policies/community-standards/hateful-conduct/.

[31] UK Department for Culture, Media and Sport and Department for Digital, Culture, Media & Sport, "Rapid Evidence Assessment (REA): The Prevalence and Impact of Online Trolling," June 26, 2019, https://www.gov.uk/government/publications/rapid-evidence-assessment-rea-the-prevalence-and-impact-of-online-trolling

[32] Meta, "Hateful Conduct," accessed May 21, 2025, https://transparency.meta.com/policies/community-standards/hateful-conduct/.

[33] Nature, "From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and LGBT communities," October 14, 2024, https://www.nature.com/articles/s41599-024-03899-1

[34] Amnesty International, "Online Violence," accessed June 10, 2025, https://www.amnesty.org/en/what-we-do/technology/online-violence/

[35] Centre for International Governance Innovation, Facebook's America-centrism Is Now Plain for All to See, October 4, 2021, https://www.cigionline.org/articles/facebooks-america-centrism-is-now-plain-for-all-to-see/

[36] The Hill, "Social media without fact-checking will destabilize the Global South," January 30, 2025, https://thehill.com/opinion/technology/5116466-social-media-without-fact-checking-will-destabilize-the-global-south/

ULTRAVIOLET