CCDH Center for Countering Digital Hate

9:41

Notifications

@anon23948909 I low key sexually harass her on IG all the time

ABUSING WOMEN IN POLITICS

How Instagram is failing women and public officials

@anon7674332 Make ~~Rape~~ Legal 🇺🇸 🇺🇸 Trump 2024 🇺🇸 🇺🇸
11.00 AM

@anon435243 ~~Ghetto ass bitch~~, and ~~pancy ass jew puppet~~ 100%
11.00 AM

@anon34599312 Hope someone leaves you for dead in a ditch. You're...
11.00 AM

@anon435243 Those who live in glass ~~wh0r3~~ houses, shouldn't through sto...
11.00 AM

CCDH Center for Countering Digital Hate

The Center for Countering Digital Hate works to stop the spread of online hate and disinformation through innovative research, public campaigns and policy advocacy.

Our mission is to protect human rights and civil liberties online.

Social media platforms have changed the way we communicate, build and maintain relationships, set social standards, and negotiate and assert our society's values. In the process, they have become safe spaces for the spread of hate, conspiracy theories and disinformation.

Social media companies erode basic human rights and civil liberties by enabling the spread of online hate and disinformation.

At CCDH, we have developed a deep understanding of the online harm landscape, showing how easily hate actors and disinformation spreaders exploit the digital platforms and search engines that promote and profit from their content.

We are fighting for better online spaces that promote truth, democracy, and are safe for all. Our goal is to increase the economic and reputational costs for the platforms that facilitate the spread of hate and disinformation.

If you appreciate this report, you can donate to CCDH at counterhate.com/donate. In the United States, Center for Countering Digital Hate Inc is a 501(c)(3) charity. In the United Kingdom, Center for Countering Digital Hate Ltd is a nonprofit company limited by guarantee.

Contents

1 Introduction

Online spaces are now the primary places where societal norms and values are negotiated and normalized, and where we learn about and discuss current events, social issues, and politics. In 2024, with democracy hanging in the balance, social media platforms are under heightened scrutiny for their role in rising polarization, stoking division, and our increasingly toxic political environment. So how are they doing? In the case of Instagram, this report finds that they may as well not be trying at all. Abuse is endemic, and there is evidence they fail to act in over 9 in 10 instances even when alerted, including comments like:

> **"We don't want blacks around us no matter who they are."**
> *Targeting Vice President Kamala Harris*
>
> **"Hope someone leaves you for a dead in a ditch."[1]**
> *Targeting Senator Marsha Blackburn*

In a diner, town hall meeting, or political rally, we would not tolerate violent, racist, or misogynistic slurs being hurled at a woman seeking to serve her community in public office. An abuser would be thrown out of the venue, fast. Yet, on Instagram, an abuser can barrage a woman with rape and death threats and can continue to use the platform with impunity.

This report is a snapshot of how platforms fail to step up to protect women and public officials. Researchers selected five Republican and five Democrat female incumbents running for office in 2024 and collected 560,000 comments on their recent Instagram posts. 1 in 25 comments – over 20,000 – were identified as likely to be "toxic" by Google's Perspective AI tool.

When our researchers reported 1000 of the worst of these comments – sexist and racist abuse, death and rape threats, and rule-breaking offenses – Instagram allowed 93% to remain on the platform. Instagram's failure to uphold and enforce its community guidelines means the platform is failing women and, by extension, our society's desire for equal opportunity and treatment for women.

The cacophony of hate speech, threats, and gendered abuse we find flooding the comment sections of prominent women politicians is united in one shared purpose: to push women out of political life.

A 2016 Data & Society study found that 41% young women between the ages of 15-29 self-censor online to avoid online harassment. In the 2020 US Congressional race, it was found that women of color candidates were more likely to receive sexist, racist, and violent abuse online. At the state level, 43% of state legislators say they have experienced threats or abuse and 40%

expressed they were unwilling to seek reelection or higher office due to abuse, according to the Brennan Center for Justice.

Algorithms reward hyper-emotive content and the engagement it generates, with amplification and visibility. Indeed, some politicians deliberately incite hatred to boost their engagement online. In our study we found that the hate in Representative Marjorie Taylor Greene's comments reflected both the type of abuse we see all women politicians in the sample endure, but also her own incitement of hate against colleagues in Congress. Just as we clearly condemn abuse against Representative Greene, there is no condoning her own cynical behavior, inciting viral cycles of abuse, much of which is itself misogynist or racist.

Countless studies have been conducted in recent years chronicling the ways women in politics face abuse online. But nothing changes if platforms refuse to act.

Instagram must enforce its existing rules against violent threats and abuse and work with experts in gender-based violence to ensure its current policies align with the lived reality of women and marginalized people in public life.

All platforms should be required to be clear on how they enforce their rules and allow visibility and scrutiny of the actions they take to address abuse. If they remove hate, tell us why. If they decide not to, again, be clear what rule it was assessed against, and why action was refused. Furthermore, platforms should report back to us – the public, regulators and advertisers – on their progress to create spaces in which women can engage without the threat of constant, persistent abuse, through the publication of regular risk assessments and progress reports.

Organizations that support women and underrepresented communities to run for office should be given the resources they need to support those targeted with online abuse, recognizing the potential for psychological and physical harm.

It is no accident that Meta and the large social media platforms that dominate our information ecosystem refuse to use their enormous resources to protect women from digital violence. The immunity gifted to social media platforms by Section 230 of the Communications Decency Act 1996 – a law passed before social media platforms even existed – means platforms can do the bare minimum to clean up their platforms, while continuing to profit lavishly. Indeed, this legal immunity has since translated into a sense of moral impunity, and, today, ironic howls of victimhood when they are criticized.

The resulting normalization of abuse, violent threats, and hate speech directed against women in public life has serious consequences to destabilize our democracies. It is our shared obligation as a society to ensure that everyone feels empowered to use their voice and participate in politics – and to hold to account those seeking to silence women.

Imran Ahmed, CEO, Center for Countering Digital Hate

CCDH Center for Countering Digital Hate

Content warning: the following report covers topics about extreme misogyny, violence against women, rape and death threats, racist hatred, and other abusive topics that some readers might find disturbing.

Readers' discretion is advised.

2 Executive Summary

<u>We collected 560,000 comments on Instagram posts from leading women politicians</u>

- We selected five Republican and five Democrat women politicians who are running for office again in 2024 and have high levels of engagement on Instagram.
  - Democrat Vice President Kamala Harris, Representatives Alexandria Ocasio-Cortez, Jasmine Crockett, Nancy Pelosi and Senator Elizabeth Warren.
  - Republican Representatives Marjorie Taylor Greene, Maria Elvira Salazar, Anna Paulina Luna, Lauren Boebert and Senator Marsha Blackburn.
- All comments in our study were posted between January 1st and June 7th, 2024.

<u>1 in 25 replies to women politicians likely to be toxic according to machine learning model</u>

- We used Google's Perspective API, a machine learning model trained to identify "toxic" text, defined as "a rude, disrespectful, or unreasonable comment".
- We found that over 1 in 25 comments is "highly likely" to be toxic, although it is unclear whether many of these comments are prohibited by Instagram's policies.

<u>Instagram fails to act on 93% of abusive comments targeting women politicians</u>

- We used Instagram's own reporting tools to report 1,000 abusive comments that targeted women politicians in our study and breach the platform's standards.
- One week after reporting the comments, Instagram had taken no action against 926 of them, equivalent to a failure to act on 93% of them. Instagram failed to act on:
  - Sexist abuse, such as "Make Rape Legal"
  - Racist abuse, like "we don't want blacks around us no matter who they are"
  - Threatening comments, like "death to her and her supporters"
- Fifth of comments came from "repeat offenders" who sent abuse more than once.

<u>Instagram is failing to act on 92% of abusive comments targeting Kamala Harris</u>

- Our research shows that Instagram fails to act on 92% of comments targeting Kamala Harris, as abuse intensifies around her presidential candidacy.
- Instagram failed to act on sexist comments claiming she is "blowjobbing to the top" and racist comments that she is "VP because you are brown and without a penis".

<u>Instagram must enforce its standards on gendered abuse</u>

- Instagram must enforce policies designed to protect women in public life.
- Organizations be equipped to support women candidates experiencing abuse.
- Lawmakers must hold social media companies accountable for failing to act.

3 Instagram claims to prohibit abusive posts and comments

Instagram's Community Guidelines claim that they "remove content that contains credible threats or hate speech" and "content that targets private individuals to degrade or shame them", yet our report suggests that they are not enforcing these guidelines and, in some cases, have decided to leave abusive comments in place even after reviewing them.[2]

Instagram prohibits attacks targeting protected characteristics including gender and race

Instagram's guidelines prohibit attacks directed at individuals based on protected characteristics, including race, ethnicity, sex and gender identity.[3]

Comments violating these guidelines include harmful stereotypes, dehumanizing speech and statements of inferiority about someone's physical appearance on the basis of their protected characteristic.[4] Instagram failed to act on comments such as "I hate this n***a" directed at Kamala Harris and "fake lashes and nasty a$$ hair" targeting Jasmine Crockett.[5]

Instagram's guidelines also prohibit statements containing severe sexualized commentary or intent to engage in sexual activity. [6] Instagram failed to act on comments violating this guideline, including "tell AOC to pop her tits out and put her sweaty toes in my mouth" and "I wanna bang squeaky", both directed at Alexandria Ocasio-Cortez.[7]

Instagram's guidelines prohibit calls for self-injury and suicide

Instagram's guidelines on bullying and harassment claim that all users are protected from calls for self-injury or suicide of a specific person on their platform.[8]

Instagram failed to act on comments in this category too, including "go fuck yourself with a rusty crowbar" and "everyone should pray to Jesus she unalives herself soon", both directed at Marjorie Taylor Greene.[9]

Instagram's guidelines prohibit comments containing threats of violence

Instagram's guidelines on violence and incitement prohibit comments containing threats, including threats of violence that could lead to serious injury or death and "glorification of gender-based violence that is either intimate partner violence or honour-based violence".[10]

Examples that Instagram failed to act on include "hope whoever attacked your husband has more people ❤❤❤❤ so they can finish the job", targeting Nancy Pelosi and "this r***** needs to be put to sleep", directed at Alexandria Ocasio-Cortez.[11]

4 We collected 560,000 comments on posts from women politicians for study

To select women politicians for our study, we created a long list of incumbent Republicans and Democrats who are running for office again in 2024. We then ranked this long list by the average number of comments each politician receives on their Instagram posts and filtered out any who had made fewer than five Instagram posts since January 1st, 2024.[12]

After this, we selected the five politicians from each party who have the highest number of average comments on their Instagram posts, resulting in the following short list for study:

- Vice President Kamala Harris (Democrat)
- Representative Alexandria Ocasio-Cortez (Democrat)
- Representative Jasmine Crockett (Democrat)
- Representative Nancy Pelosi (Democrat)
- Senator Elizabeth Warren (Democrat)
- Representative Marjorie Taylor Greene (Republican)
- Representative Maria Elvira Salazar (Republican)
- Representative Anna Paulina Luna (Republican)
- Representative Lauren Boebert (Republican)
- Senator Marsha Blackburn (Republican)

We collected 560,412 comments for analysis

Researchers compiled a list of all posts from politicians on our short list made between January 1st and June 7th, 2024. We then collected all visible comments from those posts, along with data including the date posted, the username of the poster and a URL link.[13] In total we collected 560,412 comments from 877 Instagram posts. We excluded 139,051 of these comments from our further analysis as they appeared to be replies to other users.

We used Google's Perspective API to help identify abusive comments

We used Google's Perspective API to score comments on attributes including toxicity, threat and insult.[14] We then used this data along with a keyword search to filter our full dataset of comments to those most likely to constitute a violation of Instagram's policies.[15]

Researchers assessed these comments to identify those that violate Instagram's Community Standards, arriving at a final dataset of 1,000 abusive comments.[16]

We reported 1,000 comments to Instagram to test their response

Researchers reported the 1,000 abusive comments in our final dataset to Instagram using the platform's reporting tools. After one week, researchers reviewed each comment to check if Instagram had acted to remove the comment or if it remained visible.

5 Over 1 in 25 replies to women politicians likely to be "toxic"

Our analysis of the full dataset of comments responding to women politicians in our sample shows that 1 in 25 comments is highly likely to contain "toxic" language.

Researchers performed this analysis using Google's Perspective API, a machine learning model that identifies "toxic" text, defined by Google as "a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion."[17]

This analysis was performed on the 421,361 comments in our dataset that do not mention other users, indicating they are responding to the politician whose post they were found on. We found that 17,384 of these comments were "highly likely" to be toxic, which we defined as a probability score of 0.7 or above, equivalent to over 1 in 25 of all replies.

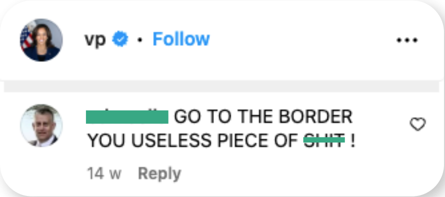
This shows that it is common for women politicians in our study to face a high rate of toxic responses from users when posting to Instagram.

Toxicity is not the same as abuse

As we outline in the previous section of this report, Instagram's policies are rightly limited to discriminatory or threatening abuse that is intended to push women and other minorities out of the public conversation. This is the right balance, providing maximum space for political speech while ensuring everyone can participate without harassment.

However, it is still the case that rising toxicity creates a polarized and hostile environment for online political discussions, making it harder for women and other minorities to play a full part in the public conversations that are essential to democracy.

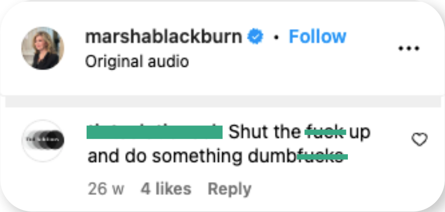
Instagram's guidelines only prohibit gendered slurs, severe sexualized commentary and threats, leaving it unclear whether the following comments rated highly for toxicity breach their standards.[18] We have therefore not included them in our analysis of Instagram's action against abusive comments that breach its policies.



vp • Follow

[redacted] GO TO THE BORDER YOU USELESS PIECE OF ~~SHIT~~ !

14 w Reply



marshablackburn • Follow
Original audio

[redacted] Shut the ~~fuck~~ up and do something dumb~~fucks~~

26 w 4 likes Reply

"You're ~~fucked~~ in the head"
Targeting Representative Jasmine Crockett

"Lay off the fucking drugs, lady. You are a global embarrassment. And btw, NOBODY remembers you in 2010 'when Clinton was president' 😂 get a grip. Your followers are dead inside."
Targeting Representative Maria Salazar

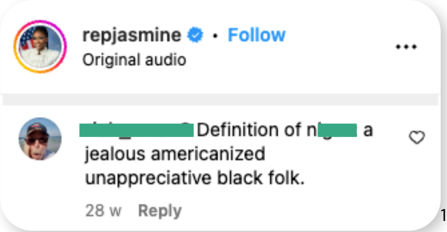
6 Instagram fails to act on 93% of abusive comments targeting women politicians

To test Instagram’s enforcement of its policies, we reported 1,000 abusive comments targeting women politicians using the platform’s own reporting tools.

We then checked to see if Instagram had hidden or removed the comments one week after reporting them, finding that 926 out of 1,000 were still visible. This amounts to Instagram’s failure to act on 93% of abusive comments targeting politicians in our study.

This helps explain why it appears some users feel that they can send abusive messages to women politicians with impunity. In many cases we found that Instagram refused to act even on comments that contained clear racist, gendered or threatening language.

Examples of gendered abuse that Instagram failed to act on are listed on the right of this page.



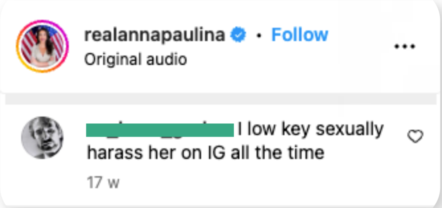
repjasmine • Follow
Original audio

[redacted] Definition of n[redacted] a jealous americanized unappreciative black folk.
28 w Reply

[19]

“We don’t want blacks around us no matter who they are.”
Targeting Vice President Kamala Harris

“AOC is [talking] like a preschooler again. I like to shove something in AOC mouth.”
Targeting Alexandria Ocasio-Cortez



realannapaulina • Follow
Original audio

[redacted] I low key sexually harass her on IG all the time
17 w

“She would sound better with a ~~fucking~~ gag in her mouth”
Targeting Representative Marjorie Taylor Greene

“Hope someone leaves you for a dead in a ditch.”[20]
Targeting Senator Marsha Blackburn

“You complete evil ~~fuking bitch,~~ the devil has a hot place in hell reserved for you”
Targeting Representative Nancy Pelosi

“Keep your ~~fucking~~ legs closed.”
Targeting Senator Elizabeth Warren

“This Lying Skank is getting Voted Out! Bye Bye Rotten Crotch!!!”
Targeting Representative Lauren Boebert

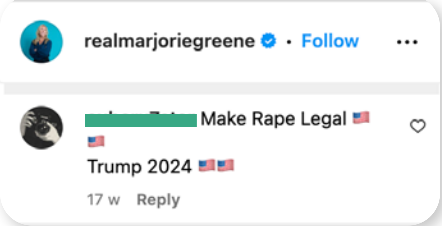
“2 stupid, lying hags”
Targeting Representative Maria Salazar

Most abusive comments targeting women politicians contain gendered terms

Analysis of our sample of abusive comments shows that most of them contain language associated with gendered abuse, such as "bitch", "rape" or "whore".

In total, 77% of the 1,000 abusive comments were categorized as containing gendered terms by at least two of our researchers. Abusive comments were included in this category if they contained gendered slurs, stereotypes about the inferiority of women, references to the physical appearance of a candidate or sexualized language.

This shows that Instagram is doing little to tackle gendered abuse in particular, despite prohibiting such abuse in its Community Guidelines.[21] This contributes to a culture of impunity where users appear to be able to send gendered abuse without fear of consequences. Examples of gendered abuse that Instagram failed to act on are listed on the right of this page.



realmarjoriegreene · Follow ···

[redacted] Make Rape Legal 🇺🇸 🇺🇸
Trump 2024 🇺🇸🇺🇸
17 w Reply

"Get raped by your president Biden"[22]
Targeting Vice President Kamala Harris

"Trump's old Ass Licking ~~Slut~~!!!"
Targeting Senator Marsha Blackburn

"Need to be back behind a bar ~~Bitch~~!!!"
Targeting Representative Alexandria Ocasio-Cortez

"~~Bitch~~, you're a paid Israeli actor. As other have pointed out, that's called treason."[23]
Targeting Representative Nancy Pelosi

"Those who live in glass ~~wh0r3~~ houses, shouldn't through stones."
Targeting Representative Lauren Boebert

"Facts??? Fact is this lying corrupt POS should be removed from her seat and sent back to the hair salon for a brain transplant"
Targeting Representative Jasmine Crockett

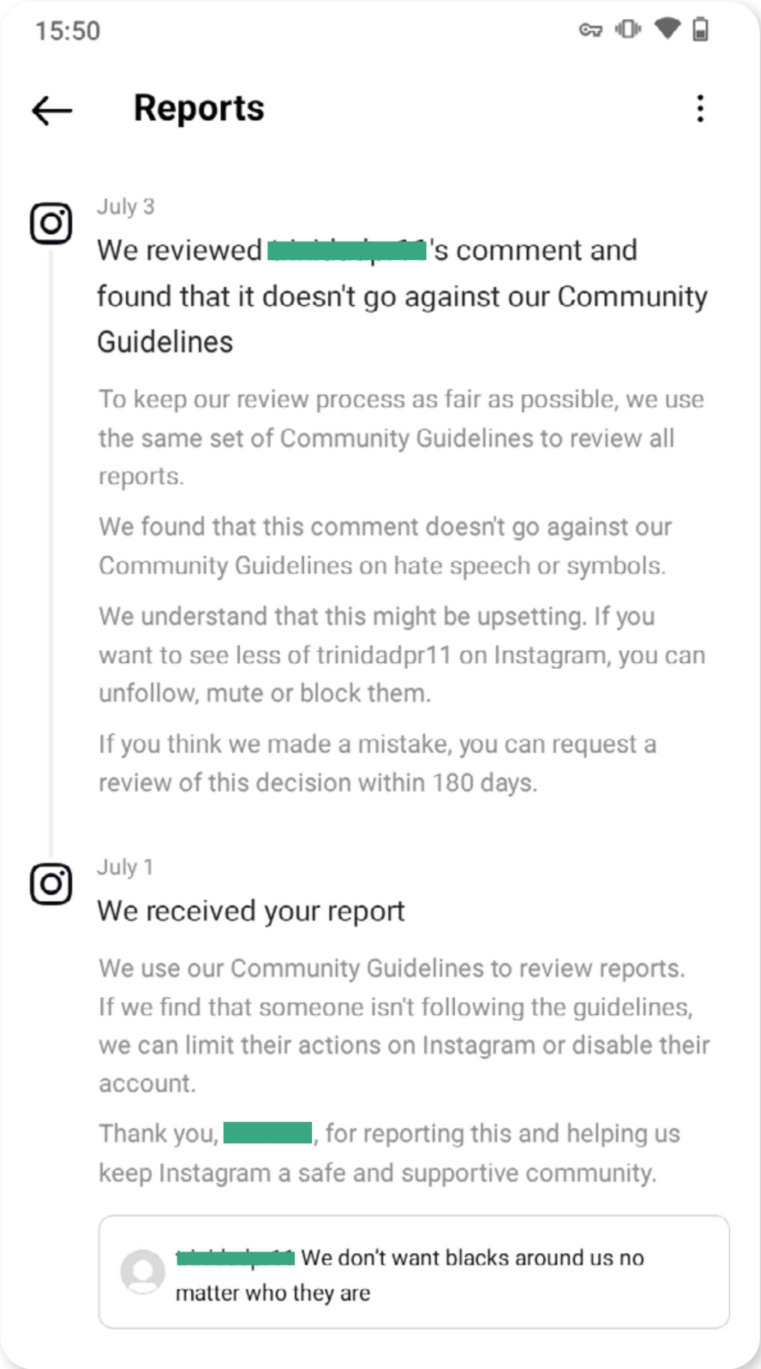
"The F@K€ C***"[24]
Targeting Representative Anna Pauline Luna

"Why do you have to lie all the time? Stupid, lying cow"
Targeting Representative Maria Salazar

"Shut it Karen. If I have no say in women's reproductive 'rights' because I'm a man, you have no say because your old ass hasn't been able to have a kid since the 90's"
Targeting Senator Elizabeth Warren

This is what it looks like when Instagram refuses to act on abuse.

One of our researchers reported this reply to Vice President Kamala Harris saying “We don’t want blacks around us no matter who they are”. They later received a notification from Instagram saying the comment “doesn’t go against our Community Guidelines”.



15:50

July 3

We reviewed [redacted]'s comment and found that it doesn't go against our Community Guidelines

To keep our review process as fair as possible, we use the same set of Community Guidelines to review all reports.

We found that this comment doesn't go against our Community Guidelines on hate speech or symbols.

We understand that this might be upsetting. If you want to see less of trinidadpr11 on Instagram, you can unfollow, mute or block them.

If you think we made a mistake, you can request a review of this decision within 180 days.

July 1

We received your report

We use our Community Guidelines to review reports. If we find that someone isn't following the guidelines, we can limit their actions on Instagram or disable their account.

Thank you, [redacted], for reporting this and helping us keep Instagram a safe and supportive community.

[redacted] We don't want blacks around us no matter who they are

Instagram fails to act on comments including calls for self-injury, suicide or death

Our analysis shows that Instagram has failed to act on comments containing violent language, including calls for self-injury, suicide or death, directed at women politicians.

These comments included calls for politicians to "die", references to "rape" or wishing violence on them.

Many of these comments violate Instagram's Community Guidelines on bullying and harassment, in which Instagram say that all users are protected from "Calls for self-injury or suicide of a specific person, or group of individuals."[25]

The few comments referencing "rape" violate Instagram's guidelines on Violence and incitement, which prohibit the "Glorification of gender-based violence that is either intimate partner violence or honour-based violence".[26]

Examples that Instagram failed to act on are listed on the right of this page.

"Hopefully you get in a fatal wreck on your way to vote."
Targeting Senator Elizabeth Warren

"They should put her body up on a cross, like the martyr she is."
Targeting Representative Alexandria Ocasio-Cortez

"Why weren't you home when old hammer head got attacked? It would have been awesome to see your knees in pieces"
Targeting Representative Nancy Pelosi

"You and your entire family deserve to die a horrible death."[27]
Targeting Vice President Kamala Harris

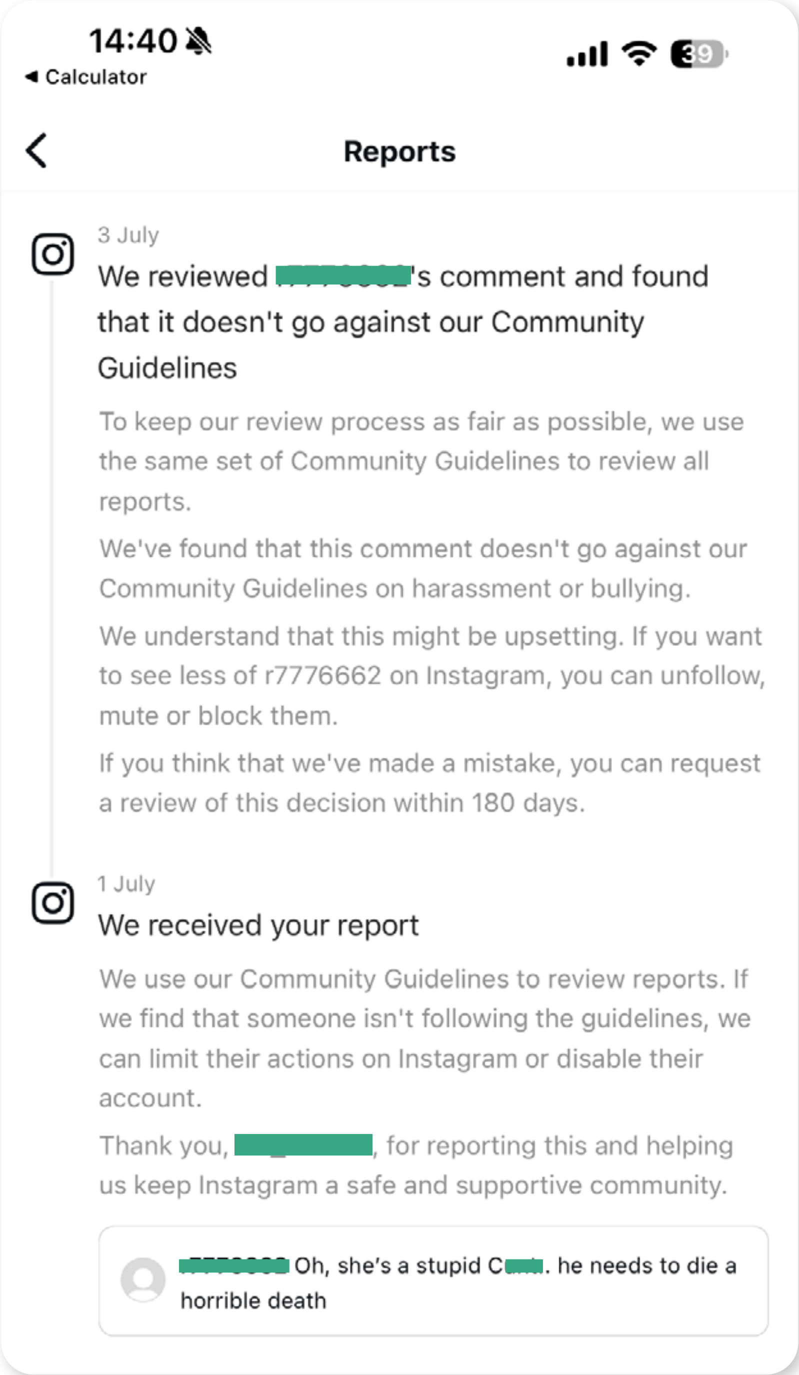
"Personally I wish that blood clot had killed her."[28]
Targeting Representative Lauren Boebert

"~~Fuck~~ MTG - death to her and her supporters"
Targeting Representative Marjorie Taylor Greene

"When I listen to you I simply want to slap you silly!! What don't you STFU and sit silently in the corner and wait for your term to end you silly Mississippi COW!
Targeting Senator Marsha Blackburn

Instagram failed to act on this comment saying she “needs to die a horrible death”.

One of our researchers received this notification after reporting a comment targeting Representative Marjorie Taylor Greene. Instagram said that the comment saying “Oh, she’s a stupid C***.. [s]he needs to die a horrible death” does not go against the platform’s Community Guidelines.[29]



3 July

We reviewed [redacted]'s comment and found that it doesn't go against our Community Guidelines

To keep our review process as fair as possible, we use the same set of Community Guidelines to review all reports.

We've found that this comment doesn't go against our Community Guidelines on harassment or bullying.

We understand that this might be upsetting. If you want to see less of r7776662 on Instagram, you can unfollow, mute or block them.

If you think that we've made a mistake, you can request a review of this decision within 180 days.

1 July

We received your report

We use our Community Guidelines to review reports. If we find that someone isn't following the guidelines, we can limit their actions on Instagram or disable their account.

Thank you, [redacted], for reporting this and helping us keep Instagram a safe and supportive community.

[redacted] Oh, she’s a stupid C[redacted]. he needs to die a horrible death

7 Fifth of abusive comments are posted by “repeat offenders”

Analysis of the abusive comments we reported to Instagram shows that over a fifth of them were posted by “repeat offenders” who had posted abuse at least twice.

In total, 221 of the 1,000 comments we reported were posted by users who had targeted the women politicians in our study during the six-month period in which we collected data.

One user had sent gendered abuse to Representative Lauren Boebert on ten occasions, referring to her as a “lying skank” and “rotten crotch”. Another sent abuse to Representative Marjorie Taylor Greene on eight occasions, repeatedly insinuating that she was in fact a man and using sexualized language.

Vice President Kamala Harris was also targeted by repeat offenders, with one commenting that “death is the answer” for black people and another repeatedly claiming that Harris’ career was the result of sexual favors.

Failure to act means abusive accounts go on to reoffend

Our previous research suggests that nearly half of accounts that platforms fail to remove for abusing women go on to reoffend.[30]

In 2022, we analyzed 235 accounts that we had previously reported to Twitter for sending abuse to high-profile women, including Vice President Kamala Harris and former presidential candidate Hillary Clinton, without Twitter acting to remove the accounts.

Two months on from Twitter’s decision to leave the accounts in place, we found that 111 of them had gone on to post further gendered abuse, equivalent to 47% of them.

This demonstrates that abusive users will go on to reoffend when platforms fail to enforce their standards, particularly if they feel enforcement is so poor they can act with impunity.

8 Instagram fails to act on 92% of abusive comments targeting Kamala Harris

Instagram failed to remove 97 out of 105 abusive comments targeting Vice President Kamala Harris, equivalent to a failure to act on 92% of abusive comments targeting her.[31]

Since the announcement that Harris is running to be the next Democratic presidential nominee, the sexist and racist abuse she faces is reported to have intensified, often branding her a "DEI hire" in an attempt to undermine her credibility.[32]

Our research indicates that social media platforms such as Instagram have not done enough to tackle racist and sexist abuse targeting Harris and other women politicians, creating a culture of impunity in which users feel they can send racist or sexist abuse without consequences. It has previously been noted that Harris receives a high rate of abuse compared even to other women politicians.[33]

Instagram failed to act on gendered abuse targeting Harris

Comments targeting Harris were littered with gendered abuse, including gender-based slurs and the common trope that women "sleep their way to the top". Instagram failed to take action on the following comments:

> "She has had more dicks in her mouth than the average gloryhole attendant does…"
>
> "madam.? duh already knew that how many guys did she blo to get where she is. oh yea n she's an un intelligent clown"
>
> "Blowjobbing to the top!"
>
> "You only care about the Jewish people because you are getting on your knees and back for a Jewish man GTFO well, millions of children and babies are getting slaughtered"
>
> "Suck enough influential and you too can be vp"

Instagram failed to act on racist comments targeting Harris' ethnicity

As the first female Vice President of Black and South Asian heritage, Harris is particularly vulnerable to racial attacks.[34] Instagram failed to remove the following comments:

> "I hate this n****"[35]
>
> "We don't want blacks around us no matter who they are"
>
> "Your mother's daughter is VP because you are brown and without a penis. She would be so proud to hear this."
>
> "What the fuk is wrong with your neck n***ooooooo??"[36]

9 Instagram fails to act on abuse encouraged by public figures

Throughout our research into abuse directed at women politicians, we took care to try and exclude comments that were directed at other public figures or Instagram users.[37]

However, it was particularly clear in the case of Representative Marjorie Taylor Greene that posts in which she targeted other public figures for criticism had inspired some Instagram users to post outright abuse as comments under her post.

Social media algorithms prioritize controversial posts that generate engagement, such as likes, share and comments.[38] When one of Greene's posts attracts abusive comments, this is likely to boost its visibility further and generate yet more engagement in a vicious cycle.

To test Instagram's enforcement of its policies against abuse triggered by a public figure, we collected a further 100 comments from posts in which Greene had attacked Democrat representative Jasmine Crockett, attorneys Fani Willis and Letitia James, Judge Arthur Engoron, and former First Lady Michelle Obama.[39]

We found that Instagram failed to take action on 96% of the abusive comments targeting other public figures under Greene's posts, including the following comments:

> "Ghetto ass bitch , and pansy ass jew puppet 100%"
> *Targeting Letitia James and Arthur Engoron*

> "All these black women trolling her should spend more time not being single mothers, raising the trash that's destroying your shitty country..."
> *Targeting Jasmine Crockett*

> "Disgusting people with no character. They should be eliminated from living in the US. Never allowed to reproduce. Take them to a black site, lock the door and never return."
> *Targeting Letitia James and Arthur Engoron*

> "We the people are fed up with this bullshit! Bring back public hanging and clean house! Enough of this shit! This is EXACTLY what the 2nd Amendment was meant for!"
> *Targeting Letitia James and Arthur Engoron*

> "She's uncouth af. Her countenance and demeanor gives shady bitch i take advantage of voters to enrich myself trading fried chicken for votes vibes."[40]
> *Targeting Fani Willis*

Notably, one of Greene's posts targeted Michelle Obama, prompting transphobic comments promoting the long-running and false conspiracy that she is a man.[41] Instagram failed to act on comments such as "All the makeup, a nice hairstyle, and Botox can never change the fact he is a man with male Chromosomes!"

10 Recommendations

The range of abuse, threats, and violence chronicled in this report seeks to damage the freedom of expression of women in politics. Countless studies in recent years have demonstrated that women political figures experience gendered abuse and toxicity on social media platforms.[42] What matters is whether platforms, with their enormous wealth of resources, have the will to act.

Political violence starts with the normalization of hateful and abusive language towards political figures. Given the dominance that social media platforms have over modern discourse, platforms like Instagram have a duty to prevent the normalization of violent and hateful political language. Based on these findings, CCDH recommends the following:

1. Instagram must transparently enforce its community guidelines against gender-based abuse and violent threats.

Instagram's failure to uphold and enforce its community guidelines means the platform is failing women and endangering their safety. Instagram must invest in more trust and safety resources to properly uphold its community rules on hateful and violent content, with special consideration to the challenges faced by political and public figures. Platforms must improve their detection of hateful, abusive, and gendered language towards women political figures.

Instagram should regularly update definitions of targeted gender-based harassment. Instagram adopted the "hidden words" feature, allowing users to block offensive words or phrases in their direct messages or comments, following CCDH's report Hidden Hate, which studied the abuse received by prominent women on Instagram via direct message.[43] However, tools like this place the onus on women and people receiving abuse to protect themselves, rather than rightfully placing the burden of responsibility on the platform to ensure abusers cannot act with impunity. Meta should invest in creating escalation channels for elected officials, especially since woman officials face extensive online abuse that cannot be mitigated by reporting comments or direct messages one by one.

Platforms must consult with subject matter experts when drafting definitions of gender-based abuse and violence and reporting incidents of gendered abuse experienced by women political figures. The rapid identification of targeted harassment campaigns is key, particularly in election periods or times of heightened scrutiny for women and marginalized people in public life.

2. Organizations helping women run for office should provide support for those facing harassment.

Racist, misogynistic, and violent threats made against women running for office not only harms candidates but can discourage those who might want to participate in democratic politics from entering the race or engaging with political debate at all. It is our collective duty to ensure that everyone is free to express themselves without fear or intimidation, which begins with detoxifying online spaces and interactions with one another. Politicians must do their part to reject sowing hatred against their colleagues in public service.

Pro-democracy organizations must recognize the impact of digital abuse on their campaigning. Organizations that support women and underrepresented communities to run for public office must be equipped with the tools to protect candidates and potential candidates from psychological and physical harm. Robust policies should be in place for employers, campaigns, and electoral organizations to protect those experiencing online harassment, including providing mental health support, anti-doxxing tools, trainings for digital safety, and budget for potential legal fees.

3. Lawmakers must act to hold social media companies accountable for failing to address abuse.

CCDH's STAR framework for social media reform emphasizes that platforms must be held to responsible standards on transparency reporting, and in this case, specifically on technology facilitated gender-based violence. Such reports should contain native analytics, top-line summaries about the quantity of content flagged by users, actions taken, user appeals, engagement with content, trust and safety spend, and other relevant data. Platforms should be required to disclose emerging trends, mitigation plans, and independent audits. To ensure compliance, social media platforms should be required to share documentation of their efforts to promote user health, well-being, and human and civil rights with regulators and the public. There must be industry-wide standards for safe platform design which prioritizes the needs of women and marginalized people in public life to prevent gendered abuse.

In the United States, lawmakers should consider including provisions against online gender-based harassment in the next reauthorization of the Violence Against Women Act. However, the largest obstacle to meaningful reform of the tech industry in the US remains Section 230.

Why did Instagram fail to remove posts that clearly violated its community rules on hate speech? The answer is because Instagram does not have to. Instagram can continue to operate with minimal safety, transparency, and accountability because of the immunity provided to it under Section 230 of the 1996 Communications Decency Act.

Instagram should not be liable for what hateful comments posted by users—but it should be liable for its behaviors as platform: for failing to act, for failing to uphold the contract it has with its users, and for ignoring user reports of repeated harassment.

Appendix: Methodology

This appendix sets out our methodology in more detail, including how we selected women politicians for the study, how we collected comments and how we analyzed them.

How we selected women politicians for our study

To select women politicians for our study, we created a long list of incumbent Republicans and Democrats who are running for office again in 2024. We then ranked this long list by the average number of comments each politician receives on their Instagram posts and filtered out any who had made fewer than five Instagram posts since January 1st, 2024.

After this, we selected the five politicians from each party who have the highest number of average comments on their Instagram posts, resulting in the following short list for study.

Name	Party	Instagram Handle
Vice President Kamala Harris	Democrat	vp
Representative Alexandria Ocasio-Cortez	Democrat	repaoc
Representative Jasmine Crockett	Democrat	repjasmine
Representative Nancy Pelosi	Democrat	speakerpelosi
Senator Elizabeth Warren	Democrat	elizabethwarren
Representative Marjorie Taylor Greene	Republican	realmarjoriegreene
Representative Maria Elvira Salazar	Republican	repmariasalazar
Representative Anna Paulina Luna	Republican	realannapaulina
Representative Lauren Boebert	Republican	repboebert
Senator Marsha Blackburn	Republican	marshablackburn

How we collected comments for analysis

Researchers collected 560,412 comments from 877 Instagram posts posted by the candidates from January 1st to June 7th, 2024, using the Google Chrome extension Comment Exporter for Instagram.[44]

We removed comments that began with an "@" symbol as this is how Instagram represents comments that are replying to another user rather than the account making the original post. Some replies did not contain this symbol and therefore could not be removed. This left a total of 421,361 comments for analysis.

How we identified comments that violate Instagram's Community Guidelines

Researchers ran Google's Perspective API model over the set of collected comments as a pre-filter.[45] This model assigns a probability score between zero and one for text for several categories, for example, the likelihood of a comment being a threat of violence. A score of zero indicates the lowest probability, while one indicates the highest.

In this case, text was rated for four categories measured by Perspective: toxicity, threat, profanity and insult.[46] Comments were included in the pre-filtered dataset if they met a pre-determined threshold in one of the four categories or if they contained a slur from the following list of keywords built using RegEx:

```
s(!|1|l)(oo|u)t|(c|k)(v|u)nt|s(l|1|!)ag|wh(o|0)r(e|3)|(b|8)(i|1|!|y)t(c|@)h|rape|shot|h(u|a)ng|h(a|@)g
```

As Perspective was applied as a pre-filter tool to collect a sample of comments for human assessment, thresholds varied. Thresholds for all candidates except Majorie Taylor Greene were as follows: Toxicity 0.6, Threat 0.4, Profanity 0.6, Insult 0.6. Thresholds for Majorie Taylor Greene were higher due to needing to process a higher total volume of comments and were as follows: Toxicity 0.85, Threat 0.85, Profanity 0.9, Insult 0.9. This left a set of 24,877 comments for human review.

To reach a final dataset of 1,000 reportable comments, researchers then manually labelled each comment in the pre-filtered dataset as reportable or not, guided by Instagram's own policies on bullying and hate speech.[47] At least two researchers looked at each comment to ensure it was reportable. For our finding on the toxicity of the dataset, researchers selected comments with a score of 0.7 or over, making them very likely to be toxic.

How we identified comments containing gendered, racist, violent or threatening language

Researchers manually labelled the 1,000 comments we reported to Instagram, checking whether they contained gendered abuse, racist abuse, violent language, or threats.

Abusive comments were labelled as containing gendered abuse if they contained gendered slurs, stereotypes about the inferiority of women, references to the physical appearance of a candidate or sexualized language.

Abusive comments were labelled as containing racist abuse if they contained a direct attack against the individual based on their race or ethnicity, including racial slurs and stereotypes about ethnic groups.

Abusive comments were labelled as containing violent language or threats if they contained comments including "threats of violence that could lead to serious injury" or death and "glorification of gender-based violence that is either intimate partner violence or honour-based violence" in accordance with some of Instagram's Community Guidelines.[48]

How we reported comments to Instagram and measured their action against them

Researchers reported the 1,000 abusive comments in our final dataset to Instagram using the platform's reporting tools between June 27th and July 24th, 2024.

One week after reporting, researchers reviewed each comment to check if Instagram had acted to remove the comment or if it remained visible.

Appendix: Limitations

This appendix sets out some of the limitations of our report, including the inability to collect every comment from the Instagram posts we studied and that our use of Google's Perspective API underestimated the true scale of abusive content.

Limitations of Comment Exporter for Instagram

We used Comment Exporter for Instagram to capture up to 10,000 text comments for each of the 877 posts in our study, exporting them as CSV files. The tool does capture nested comments in which users reply to one another and captures emojis as Unicode characters, but does not capture embedded media such as images and GIFs.

In some cases, there was a discrepancy between the total number of comments on the post and the total number of comments exported as the tool does not capture comments from private users. Based on our testing, we estimate that 79.24% of all comments are visible and accessible for data collection

Comment Exporter does not indicate if a comment is a reply to the original post or another user. Researchers removed comments that began with usernames preceded by the "@" symbol, but some user-to-user replies may have remained in the larger set. All such user-to-user replies were manually removed from the smaller set of reported comments.

Limitations of Google's Perspective API

Like all machine learning models, Google's Perspective API is not perfectly accurate, although it has performed well on tests historically.[49] In addition, it is prone to bias and as such might label text as highly likely to be toxic or vice versa in error.[50]

We primarily used Perspective to pre-filter our dataset to the most toxic comments, and therefore those that were most likely to violate Instagram's Community Guidelines. All comments were checked by researchers before being reported to Instagram, mitigating the risk of false positives.

Endnotes

[1] The full version of this comment is “Hope someone leaves you for a dead in a ditch. You’re a disgusting disgrace to this earth; I don’t even live in the United States but if you are the future of America, god help you all! You should be so disgustingly ashamed for what you stand for and what you practice. Being a ‘Christian woman’ and standing against BASIC HUMAN RIGHTS is despicable. You should be put out and put to shame.”

[2] “We want to foster a positive, diverse community. We remove content that contains credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages. We do generally allow stronger conversation around people who are featured in the news or have a large public audience due to their profession or chosen activities.”

“Community Guidelines”, Instagram, accessed 25 July 2024, https://help.instagram.com/477434105621119

[3] “We define hate speech as a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.”

“Hate Speech”, Meta, accessed 27 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/

[4] “We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate or exclude specific groups, and that are often linked with offline violence.”

“Hate Speech”, Meta, accessed 27 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/

[5] The first comment in this sentence has been edited to mask a racial slur.

[6] “Everyone is protected from: Statements of intent to engage in a sexual activity or advocating to engage in a sexual activity; Severe sexualized commentary.”

“Bullying and harassment”, Meta, accessed 24 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/bullying-harassment/

[7] The first comment in full is “I m totally uninterested in your message about climate change. Tell AOC to pop her tits out and put her sweaty toes in my mouth. Inshallah, it will happen!”.

[8] “Everyone is protected from: Calls for self-injury or suicide of a specific person, or group of individuals.”

“Bullying and harassment”, Meta, accessed 24 July 2024, https://transparency.meta.com/en-gb/policies/

community-standards/bullying-harassment/

[9] The second comment in full is “I hope she unalive herself. Even better if someone else doesn’t because this thing, I can’t even call it a woman after all the sex changes it’s had done. Everyone should pray to Jesus she unalives herself soon.”

[10] “Everyone is protected from the following threats: Threats of violence that could lead to death (or other forms of high-severity violence); Threats of violence that could lead to serious injury (mid-severity violence).”

“In addition to all of the protections listed above, we remove the following: Glorification of gender-based violence that is either intimate partner violence or honor-based violence”

“Violence and incitement”, Meta, accessed 23 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fcredible_violence

[11] The first comment in full is “I hope whoever attacked your husband has more people ❤️❤️❤️❤️ so they can finish the job ❤️❤️❤️ its what u deserve ❤️❤️❤️ soulless hack”. The second comment has been edited to mask a slur, and in full is “Please. This r***** needs to be put to sleep. That is stupid on a colossal level.”

[12] Social Blade, https://socialblade.com/

[13] According to testing carried out by our researchers, we estimate that 79.24% of all comments are visible and accessible for data collection. For more detail see Appendix: Methodology.

[14] “How it Works”, Perspective API, accessed 18 July 2024, https://perspectiveapi.com/how-it-works/

[15] See Appendix: Methodology for keyword search

[16] “Bullying and harassment”, Meta, accessed 27 June 2024, https://transparency.meta.com/en-gb/policies/community-standards/bullying-harassment/

“Hate speech”, Meta, accessed 27 June 2024, https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/

“Violence and incitement”, Meta, accessed 27 June 2024, https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/

[17] “How it Works”, Perspective API, accessed 18 July 2024, https://perspectiveapi.com/how-it-works/
See Appendix: Methodology for more details.

[18] ”Hate speech“, Meta, accessed 25 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fhate_speech

“Bullying and harassment”, Meta, accessed 25 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/bullying-harassment/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fbullying

19 This comment has been edited to mask a racial slur.

20 The full version of this comment is “Hope someone leaves you for a dead in a ditch. You’re a disgusting disgrace to this earth; I don’t even live in the United States but if you are the future of America, god help you all! You should be so disgustingly ashamed for what you stand for and what you practice. Being a ‘Christian woman’ and standing against BASIC HUMAN RIGHTS is despicable. You should be put out and put to shame.”

21 “We define hate speech as a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation.”

“Hate speech”, Meta, accessed 25 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fhate_speech

22 The full version of this comment is “Your CIA garbage trespasser king is reigning in my daughter Juan Xu homescreen of her two TV sets to LOL at your lamest power get raped by your president Biden and bow your Virginia to your end of JEHOVAH’S CURSE. Amen.”

23 The full version of this comment is “This asshat @speakerpelosi said pro-Palestinian people are paid by Russians. Bitch, you’re a paid Israeli actor. As others have pointed out, that’s called treason”.

24 This comment has been edited to mask a gendered slur.

25 “Everyone is protected from: Calls for self-injury or suicide of a specific person, or group of individuals.”

“Bullying and harassment”, Meta, accessed 24 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/bullying-harassment/

26 “In addition to all of the protections listed above, we remove the following: Glorification of gender-based violence that is either intimate partner violence or honour-based violence”

“Violence and incitement”, Meta, accessed 27 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/

27 The full version of this comment is “You and your entire family deserve to die a horrible death. And the Almighty Allah will deal with you. You fk trist peace of shiiiit! You belong in prison with ntYahoo.”

28 The full version of this comment is “Personally I wish that blood clot had killed her. What makes anyone think that we should pay her medical bills? Her stupid son is a criminal , must have learned that behavior from home. It’s not all her retarded ex husband Gay-sons fault!”

29 The comment in this sentence has been edited to mask a gendered slur.

30 “Nearly half of the accounts Twitter fails to remove for abusing women go on reoffend”, Center for

Countering Digital Hate, 12 January 2022, https://counterhate.com/blog/nearly-half-of-the-accounts-twitter-fails-to-remove-for-abusing-women-go-on-reoffend/

31 "Kamala Harris is officially running for president. Will any Democrats challenge her?" NBC News, 21 July 2024, https://www.nbcnews.com/politics/2024-election/kamala-harris-running-president-democratic-challengers-vp-biden-rcna162939

32 "'Sexist' falsehoods target Kamala Harris after Biden drops out", France24, 23 July 2024, https://www.france24.com/en/live-news/20240722-sexist-falsehoods-target-kamala-harris-after-biden-drops-out

"Kamala Harris faces racial 'DEI' attacks amid campaign for the 2024 presidency", ABC News, 24 July 2024, https://abcnews.go.com/Politics/kamala-harris-faces-racial-dei-attacks-amid-campaign/story?id=112196464

33 "Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online", 25 January 2021, Wilson Center, https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online

34 Harris could become the first female president after years of breaking racial and gender barriers, ABC News, 21 July 2024, https://abcnews.go.com/US/wireStory/harris-endorsed-biden-become-woman-black-person-president-112146028

35 This comment has been edited to mask a racial slur.

36 This comment has been edited to mask a racial slur.

37 Researchers manually checked the set of 1,000 comments to ensure comments were directed at the politician who uploaded the post. See Appendix: Data Collection for information on removing replies to other Instagram users.

38 "Shedding More Light on How Instagram Works", Instagram, 8 June 2021, https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works

"Deep dive into Meta's algorithms shows that America's political polarization has no easy fix", AP News, 28 July 2023, https://apnews.com/article/facebook-instagram-polarization-misinformation-social-media-f0628066301356d70ad2eda2551ed260

39 realmarjoriegreene, Instagram, 24 May 2024, https://www.instagram.com/p/C7W0LyCt0QF

realmarjoriegreene, Instagram, 15 February 2024, https://www.instagram.com/p/C3Yc0WQuN1c

realmarjoriegreene, Instagram, 21 March 2024, https://www.instagram.com/p/C4yweIdunK2

realmarjoriegreene, Instagram, 9 January 2024, https://www.instagram.com/p/C14aYz1Ob_8

40 The full version of this comment is "She's uncouth af. Her countenance and demeanor gives shady bitch i take advantage of voters to enrich myself trading fried chicken for votes vibes. ▯She lacks the solemnity her

position requires."

41 "False story about former first lady's mother originated as satire", AP News, 22 July 2022, https://apnews.com/article/fact-check-michelle-obama-mother-satire-623260875576

42 "An Unrepresentative Democracy: How Disinformation and Online Abuse Hinder Women of Color Political Candidates in the United States", Center for Democracy and Technology, 27 October 2022, https://cdt.org/insights/an-unrepresentative-democracy-how-disinformation-and-online-abuse-hinder-women-of-color-political-candidates-in-the-united-states/
"Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online", 25 January 2021, Wilson Center, https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online

"Digital microaggressions and everyday othering: an analysis of tweets sent to women members of Parliament in the UK", 4 January 2021, https://doi.org/10.1080/1369118X.2021.1962941
"Tackling Online Abuse and Disinformation Targeting Women in Politics". Carnegie Endowment for International Peace", 20 November 2020, https://carnegieendowment.org/research/2020/11/tackling-online-abuse-and-disinformation-targeting-women-in-politics?lang=en

43 "Hide comments or message requests that you don't want to see on Instagram", Help Center, Instagram, accessed 2 August 2024, https://www.facebook.com/help/instagram/700284123459336
"CCDH made Instagram act on violence against women and girls", Center for Countering Digital Hate, 30 January 2024, https://counterhate.com/blog/ccdh-made-instagram-act-on-violence-against-women-and-girls/

44 "Comments Exporter", WeBooster, accessed 12 June 2024, https://chromewebstore.google.com/detail/cckachhlpdnncmhlhaepfcmmhadmpbgp

45 "Perspective API", Jigsaw, accessed 12 June 2024, https://perspectiveapi.com/

46 "Attributes & Languages", Jigsaw, accessed 25 June 2024, https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

47 "Bullying and harassment", Meta, accessed 27 June 2024, https://transparency.meta.com/en-gb/policies/community-standards/bullying-harassment/

"Hate speech", Meta, accessed 27 June 2024, https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/

"Violence and incitement", Meta accessed 27 June 2024, https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/

48 "Violence and incitement", Meta, accessed 23 July 2024, https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fcredible_violence

[49] “Model Cards”, Jigsaw, accessed 25 June 2024, https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US&tabset-20254=3

[50] “Better Discussions with Imperfect Machine Learning Models”, Jigsaw, accessed 24 July 2024, https://medium.com/jigsaw/better-discussions-with-imperfect-models-91558235d442

CCDH Center for Countering Digital Hate

Abusing Women in Politics:
How Instagram is failing women and public officials

Published August 2024