

# SAFETY OFF

How threats and abuse surged following Meta's policy changes

👍👎👤 112

81 comments 34 shares

👍 Like

💬 Comment

➦ Share



I'm praying you



Sar  
Hopefully someone



I hope you





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

Our mission is to protect human rights and civil liberties online. Social media and AI 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 environments where harmful and dangerous content can spread rapidly, often without meaningful oversight or accountability.

Social media companies erode basic human rights and civil liberties by enabling the amplification of harmful content and distorting the online information ecosystem.

**If you appreciate this report, you can donate to CCDH at [counterhate.com/donate](https://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 non-profit company limited by guarantee.**

# CONTENTS

---

|                                              |    |
|----------------------------------------------|----|
| 1. Introduction                              | 4  |
| 2. Key Findings                              | 7  |
| 3. Studying Abuse of Politicians on Facebook | 9  |
| 4. Abuse Surged After Policy Changes         | 10 |
| 4.1. Violence and Incitement Surged          | 11 |
| 4.2. Hateful Conduct Surged                  | 12 |
| 4.3. Bullying and Harassment Surged          | 13 |
| 5. Threats Against Trump Surged              | 14 |
| 6. Meta Cut Enforcement                      | 15 |
| Appendix: Methodology                        | 16 |
| Appendix: List of Members Studied            | 21 |
| Appendix: Statistical Significance           | 22 |
| Appendix: Share of Abusive Comments          | 24 |
| References                                   | 26 |

# 1. INTRODUCTION

---

In April, as we were completing this report, a man forced his way through the security screening at the White House Correspondents' Dinner in Washington, D.C., carrying a shotgun, a pistol, and several knives.<sup>1</sup>

The President was rushed from the room. A Secret Service officer was shot, saved only by his body armor.<sup>2</sup> It was the third attempt on Donald Trump's life in under two years.

The attack arrived in a country that has nearly stopped being shocked by such things. The firebombing of Governor Josh Shapiro's residence while his family slept.<sup>3</sup> The murder of Minnesota House Speaker Melissa Hortman and her husband Mark, shot in their own home.<sup>4</sup> The killing of Charlie Kirk on a Utah campus, in front of a crowd.<sup>5</sup> From the President to first-term members of Congress, from the left to the right, public service in America now carries the risk of violence.



Away from the headlines, a quieter retreat is under way. Lawmakers are canceling town halls or moving them online.<sup>6</sup> Election officials are leaving the job.<sup>7</sup> Representatives admit, in private, that the fear of being targeted now shapes how they vote.<sup>8</sup>

Meta once understood the danger in this. For as long as it has had rules, it has had rules against intimidation and harassment, because, in Mark Zuckerberg's own words in 2019, such conduct "can have the effect of restricting others' right to speak."<sup>9</sup> That was the principle. You protect speech by protecting people from being frightened into silence.

In January 2025, Meta abandoned it. Under the banner of "More Speech and Fewer Mistakes," the company rolled back the automated systems it had used to find and remove violent threats, hate, and harassment.<sup>10</sup> At the precise moment that political intimidation was becoming a democratic emergency, Meta switched off much of the machinery built to contain it. This was not an accident of policy. It was a business decision, a choice to spend less on keeping people safe, and its consequences were entirely foreseeable.

This report measures them. Drawing on an analysis of nearly 8 million Facebook comments directed at Members of Congress, we find that abuse has tripled since Meta's changes:

- ▶ Violent threats quadrupled, from 1,800 in the six months before the changes to 7,600 in the six months after.<sup>11</sup>
- ▶ Hate quadrupled, from 6,900 to 30,000.<sup>12</sup>
- ▶ Harassment doubled, from 15,700 to 39,900.<sup>13</sup>

Threats against President Trump doubled during the same period.<sup>14</sup> Many of these threats could qualify as felony offenses.

As abuse rose, Meta's enforcement fell. The company's own transparency reports show that it cut its proactive enforcement of this kind of content by roughly half in the months following its policy changes.<sup>15</sup> The surge in abuse and the collapse in enforcement track one another almost exactly.

None of this should surprise anyone, least of all Meta. In our earlier report, *More Transparency and Less Spin*, we used the company's own past statements and data to predict that these changes would reduce enforcement and that abuse would climb.<sup>16</sup> Both have now happened.

We are not alone in reaching this conclusion. Responding to the findings in our report, Senator John Curtis (R-Utah), a member of the influential U.S. Senate Committee on Commerce, Science, and Transportation that has oversight over issues related to platform liability, put it plainly: "When companies reduce oversight in areas like violence, hate, and harassment, it should not be any surprise to see those harms increase."

I do not write about this as an abstraction. Ten years ago this month, my friend and colleague Jo Cox, a Member of the British Parliament, was murdered by a man who had been radicalized, in part, online. As this report goes to press, we are only days away from the tenth anniversary of her death. In the weeks after her death, I came to understand something simple and frightening: the companies whose platforms help turn private grievance into public violence carry almost no obligation to prevent it. CCDH exists because of that realization. Ten years later, the platforms are bigger, the harms are better documented, and the obligation is still close to nonexistent.

It does not have to stay that way. Our history offers countless examples of powerful new industries outpacing the public's ability to hold them accountable. In the age of the railroads and the trusts, we did not abolish new industries. We built rules. We can do the same here. Meta should restore proper enforcement of its policies on violent threats, hate, and harassment, and be far more open about what it permits, what it removes, and why. Users have a right to know the terms under which their public square is governed. Transparency and legal checks and balances that comply with the Constitution are a potent means of balancing private interests with the public good. We should not balk at doing the same with social media.

When intimidation becomes the price of public service, fewer people will be willing to pay it. We can already see who is being driven out and who is being deterred from ever stepping forward. A democracy that cannot keep safe the people who serve it is a democracy that has quietly begun to hollow out. Meta knew this once. They should remember it again, before it is too late.

**Imran Ahmed**  
**CEO, Center for Countering Digital Hate ●**

## EXPERT AND LAWMAKER REACTIONS

The findings in this report have prompted concern from lawmakers and civil society organizations at the forefront of addressing political violence and online harms. Their perspectives include Speaker Emerita Nancy Pelosi (D-CA), whose family was directly affected by political violence; Senators Adam Schiff (D-CA) and John Curtis (R-UT) and Representative Jamie Raskin (D-MD), who serve in influential roles on congressional committees with oversight of the digital ecosystem and online platforms; and South Carolina State Representative Brandon Guffey (R-District 48), whose son died in a social media-enabled sextortion scam.

The report also includes responses from the Public Service Alliance, which works to protect and support public servants facing escalating threats and harassment, and Everytown for Gun Safety, one of the nation's leading organizations working to prevent gun violence. Together, their statements underscore the profound societal consequences of online abuse and the growing imperative for greater platform accountability.



### **U.S. Senator John Curtis (R-UT):**

"When companies reduce oversight in areas like violence, hate, and harassment, it should not be any surprise to see those harms increase. Similarly, the reported surge in abusive and threatening content directed at public officials is deeply concerning, particularly in light of recent events. Ultimately, this report seriously questions the ability of companies to self-regulate and underscores the need for Congressional action to ensure accountability and protect the rights and safety of users."

### **U.S. Senator Adam Schiff (D-CA):**

"The surge in hate on Meta's platforms enumerated in this report should concern every American. By rolling back their platforms' harassment and intimidation rules last year, Meta has allowed false and dangerous content to proliferate — making it harder for our country to bridge the political divide and increasing the threat of political violence. It's time for social media companies to fulfill their responsibility to crack down on harmful content, which can have dangerous off-screen consequences."

### **Speaker Emerita Nancy Pelosi (D-CA):**

"Public service should never come with a torrent of threats and abuse. Reports showing the scale of harassment on Meta's platforms are deeply troubling, especially in light of decisions to weaken content moderation. When Meta's business model is one that profits from hateful engagement, they have a responsibility to prevent that engagement from devolving into intimidation or violence. If harassment becomes the cost of participation, it risks driving good people away from serving their communities. This is especially true for mothers, who avoid participating in our democracy to spare their children the trauma of hearing such vitriol. We must continue working to strengthen accountability, protect public servants and uphold the integrity of our discourse."

**U.S. Rep. Jamie Raskin (D-MD):**

"Different social media entities need to decide whether they want to platform hate, harassment and violent threats. This report shows that the rules matter in controlling how much pollution of the moral and social environment we are going to suffer and tolerate."

**South Carolina State House Rep. Brandon Guffey (R-District 48):**

"When online threats and harassment triple, that's not simply a moderation issue. It's a public safety issue. We can defend the First Amendment while also expecting technology companies to take credible threats, incitement, and targeted abuse seriously. No one should have to choose between online spaces and their safety as a citizen. Especially children."

**Isabella Ulloa, Founder & CEO, Public Service Alliance (PSA):**

"Violent threats against America's public servants and their families — including doxxing and other online threats — have been skyrocketing for over a decade, often fueled by lack of responsible governance across social media platforms. CCDH's latest report is a stark reminder of this reality. Curbing the proliferation of online threats and protecting First Amendment rights are not mutually exclusive, and failing to act is a choice with real-world consequences that ultimately make America weaker."

**Nick Suplina, SVP of Law and Policy, Everytown for Gun Safety:**

"This report confirms a chilling reality: when tech giants like Meta roll back safety standards, they hand a megaphone to hate and invite political violence into the digital public square. The quadrupling of violent threats against elected officials is a predictable consequence of Meta's decision to reduce content moderation. The distance between online vitriol and real-life gun violence is dangerously short. We cannot have a healthy democracy if the price of public service is a barrage of death threats. Meta must stop prioritizing engagement over safety and immediately reinstate these vital safeguards."

## 2. KEY FINDINGS

### **META CUT ENFORCEMENT OF ITS POLICIES ON ABUSE**

- ▶ On January 7, 2025 Meta announced policy changes to roll back 'proactive enforcement' for abuse, including the three following policy areas:
  - ▷ Violence and incitement
  - ▷ Hateful conduct
  - ▷ Bullying and harassment<sup>19</sup>
- ▶ Meta's own transparency reports show that it halved proactive enforcement of abuse on Facebook in the six months following its policy changes.<sup>20</sup>

### **WE ANALYZED ABUSE DIRECTED AT MEMBERS OF CONGRESS**

- ▶ We collected nearly 8 million comments on Facebook posts made by Members of Congress in the six months before and after Meta's policy changes.
- ▶ Using AI, we measured the volume of comments targeting Members of Congress with abuse that violate Meta's policies before and after the policy changes.

### **ABUSE SURGED AFTER META'S POLICY CHANGES**

- ▶ Abusive comments tripled after Meta's policy changes, rising from 24,000 abusive comments in the six months before to 78,000 in the six months after.<sup>21</sup>
  - ▷ Violence and incitement quadrupled, including comments such as:
    - ▷ "Hopefully someone blows your brains out asap."
    - ▷ "Rope and tree for you !! Treason."
  - ▷ Hateful conduct quadrupled, including comments such as:
    - ▷ "Worthless bitch should be picking cotton maybe then she'd have value."
    - ▷ "We have credible evidence that Raskin is all about the Benjamins"
  - ▷ Bullying and harassment doubled, including comments such as:
    - ▷ "If you're so concerned about this country, you can best serve it by committing suicide"
    - ▷ "I'm praying you get cancer"
- ▶ This surge in abuse following Meta's policy changes is statistically significant.

### **THREATS AGAINST DONALD TRUMP SURGED AFTER POLICY CHANGES**

- ▶ Comments targeting President Donald J. Trump with violence and incitement doubled after Meta's policy changes. Example comments include:
  - ▷ "4 inches to the left and the shooter would have Made America Great Again."
  - ▷ "Trump deserves a bullet through his head."

### 3. STUDYING ABUSE OF POLITICIANS ON FACEBOOK

We used AI to analyze 8 million comments on Facebook posts from politicians

To measure the impact of Meta’s policy changes, we used AI to analyze nearly 8 million comments on Facebook posts from 100 Members of the House of Representatives.

#### HOW WE SELECTED POLITICIANS TO STUDY

Researchers created a list of Representatives and matched them to Facebook pages that they operate. From this list, they selected 50 Republicans and 50 Democrats with the most followers to carry forward into post and comment collection.

#### HOW WE COLLECTED FACEBOOK COMMENTS

Researchers first collected all Facebook posts made by Representatives in our analysis in the six months before and after Meta’s policy changes were announced on January 7, 2025.<sup>22</sup> Where Representatives had multiple pages, data was collected from all of them. This resulted in a set of Facebook posts from which researchers collected a total of 7,868,763 comments to take forward into analysis to identify abusive comments.

#### HOW WE IDENTIFIED ABUSIVE COMMENTS

This report identifies comments as “abusive” if they violate Meta’s current policies in three areas: violence and incitement, hateful conduct, or bullying and harassment.<sup>23</sup>

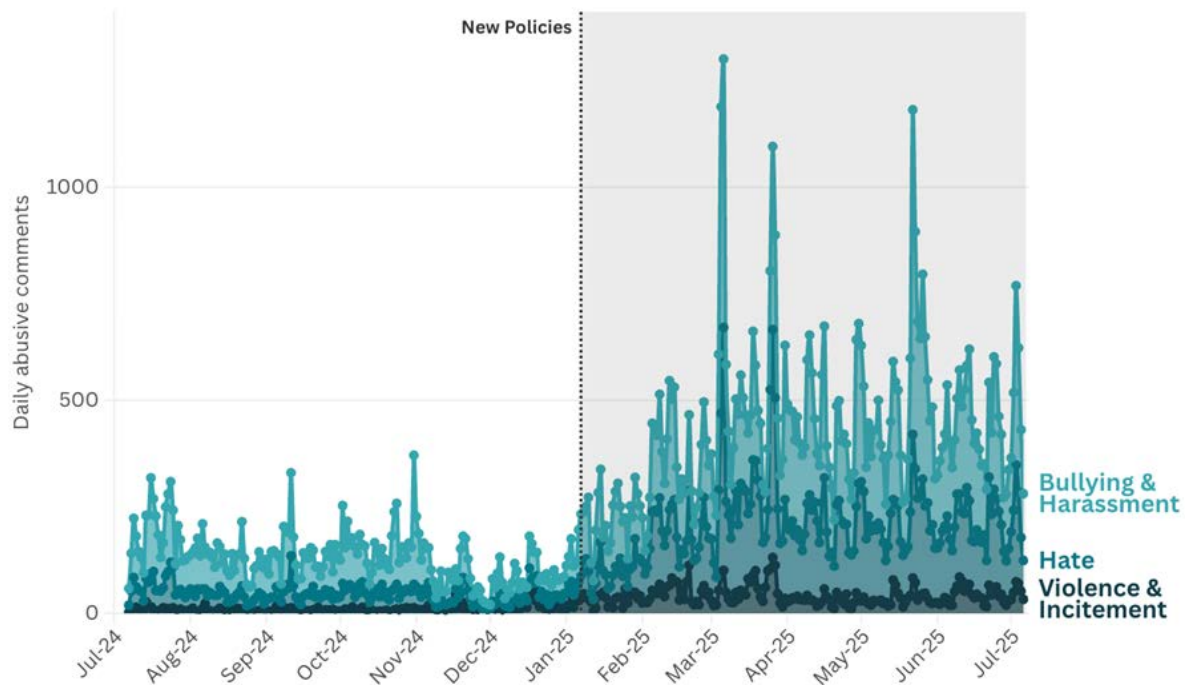
Analysis was carried out using an AI system developed to identify comments in our dataset that are likely to violate each of the three policy areas outlined above. This enabled us to measure the volume of abusive comments in our dataset over time.

Researchers manually assessed the AI system’s performance in identifying abuse in samples of comments from our dataset, finding that the system functioned reliably with an average F1 accuracy score of 79% across all three policy areas studied.<sup>24</sup>



## 4. ABUSE SURGED AFTER POLICY CHANGES

### Abusive comments targeting politicians tripled



We used AI to analyze nearly 8 million Facebook comments directed at Members of Congress to investigate whether Meta’s policy changes led to a surge in abuse.

Our analysis finds that the volume of abusive comments was three times higher in the six months after Meta’s policy changes than it was in the six months prior.

Comments were identified as abusive if they violated Meta’s current policies in three areas: violence and incitement, hateful conduct, or bullying and harassment.<sup>25</sup>

Our analysis identified an estimated 102,000 abusive comments targeting politicians, of which 24,000 were posted before Meta’s policy changes and 78,000 were posted after. This represents an overall tripling of abusive comments.<sup>26</sup>

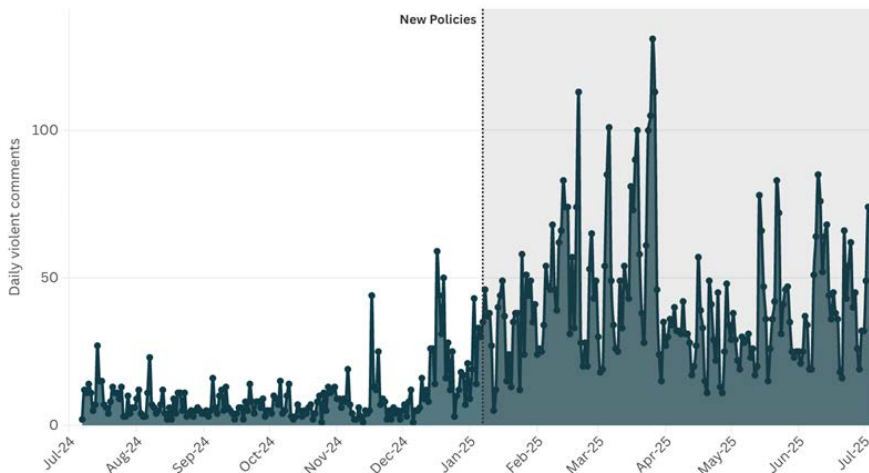
While abuse sometimes spiked around specific events or viral trends, our analysis finds a statistically significant step-change in total volume after the policy changes. We also found a shift from a slight downwards trend in abuse before the changes to a slight upwards trend after. More details on this supplementary analysis are available in Appendix: Statistical Significance.

These findings are consistent with the hypothesis that Meta’s policy changes have contributed to a surge in the volume of abuse directed at politicians on Facebook.

The following sections of the report examine how the volume of abuse surged in all three categories of abuse under analysis and was accompanied by a rise in threats directed at the President Donald J. Trump.

## 4.1 VIOLENCE AND INCITEMENT SURGED

### Comments violating Meta's violence and incitement policy quadrupled



In the previous section, we found that various kinds of abuse surged in the wake of Meta's policy changes.

We identify a similar surge when examining just those comments that violate Meta's policies on violence and incitement, which were four times higher in the six months after Meta's policy changes than in the six months prior.

Meta's current policies on violence and incitement prohibit content expressing an intention, aspiration, or call for violence against a target.<sup>27</sup> The policy states that these may take various forms including "statements of intent, calls for action, advocacy, expressions of hope, aspirational statements and conditional statements."<sup>28</sup>

Our analysis identified an estimated 9,400 comments violating these policies, of which 1,800 were posted before Meta's policy changes and 7,600 were posted after. This represents an overall quadrupling of violent comments.<sup>29</sup>

Comments in this category include calls for the murder, assault, rape as well as more general calls for politicians to be killed or die.

"This little p\*\*\*\*, and I mean, little p\*\*\*\*, he's that f\*\*\*\*\* dead little short,m\*\*\*\*\*.He's gonna f\*\*\*\*\* die screaming this m\*\*\*\*\*"

– Targeting Mike Johnson

"I hope you get shot."

– Targeting Nancy Mace

"You should have lost both eyes you traitorous rat."

– Targeting Dan Crenshaw

"You better watch it Hakeem, you'd be a quick and easy KO little man, I don't care how much security you have BOY 🙌"

– Targeting Hakeem Jeffries

"Rope and tree for you !! Treason"

– Targeting Robert Garcia

"They need to gut you and send you back to Somalia."

– Targeting Ilhan Omar

"Hopefully someone blows your brains out asap"

– Targeting Marjorie Taylor Greene

"Every real American hero and patriot should 💀 at least 10-20 MAGA criminals and then receive the "Medal of Honor" for his patriotic act against the evil and enemy of America to save democracy and freedom"

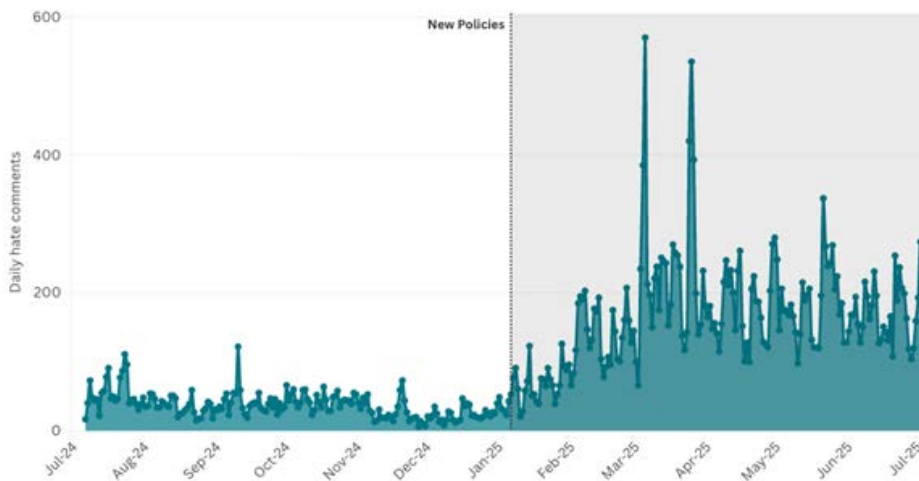
– Targeting Marjorie Taylor Greene

"Byron Donalds... you will hang from the nearest tree for betraying the human race"

– Targeting Byron Donalds

## 4.2 HATEFUL CONDUCT SURGED

Comments violating Meta’s hateful conduct policy quadrupled



We find that comments violating Meta’s policies on hateful conduct were four times higher in the six months after Meta’s policy changes than in the six months prior.

Meta’s current policies on hateful conduct prohibit attacks based on characteristics such as race, sexual orientation or sex. Attacks can take the form of calls for harm, dehumanization, harmful stereotypes and slurs.<sup>30</sup>

Mark Zuckerberg has previously defended such policies, saying that “we know from history that dehumanizing people is the first step towards inciting violence.”<sup>31</sup>

Our analysis identified an estimated 36,900 comments violating these policies, of which 6,900 were posted before Meta’s policy changes and 30,000 were posted after. This represents an overall quadrupling of hateful comments.<sup>32</sup>

Comments in this category target politicians with racist, sexist and homophobic slurs, as well as harmful stereotypes about race, religion, and other characteristics. There were also a number of calls to exclude women from political office entirely.

“Worthless bitch should be picking cotton maybe then she’d have value”  
– Targeting Jasmine Crockett

“Good job but no women should be In Control over a state people anywhere”  
– Targeting Nancy Mace

“Too much curry in your tikka masala? Stop spreading lies paki”  
– Targeting Pramila Jayapal

“What a lying sack of \$hit this Muslim bit(h is. Look it up”  
– Targeting Pramila Jayapal

“We have credible evidence that Raskin is all about the Benjamins.”  
– Targeting Jamie Raskin

“Dude sounds like a straight up creep. But considering he’s part of the pedo Christian religion, it’s not surprising.”  
– Targeting Mike Johnson

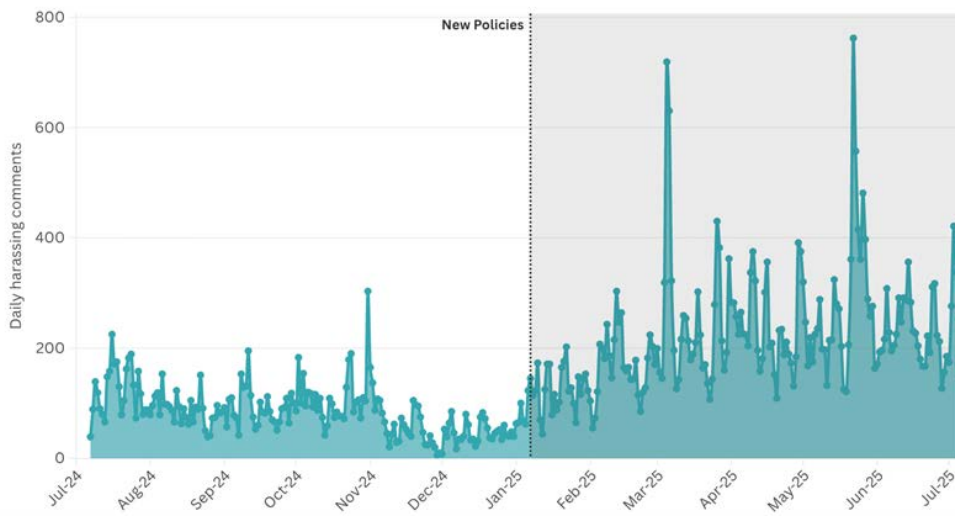
“No you did n\*\*\*\*\*”  
– Targeting Wesley Hunt

“She is a fucktard and a dike”  
– Targeting Becca Balint

“Go home beaner”  
– Targeting Jimmy Gomez

## 4.3 BULLYING AND HARASSMENT SURGED

Comments violating Meta's bullying and harassment policy doubled



We find that comments violating Meta's policies on bullying and harassment were over twice as high in the six months after Meta's policy changes than in the six months prior.

Meta's current policies on bullying and harassment prohibit calls for someone to commit suicide, various forms of sexual harassment, and threatening to release private details such as home addresses (a practice known as "doxxing").<sup>33</sup>

Our analysis identified an estimated 55,600 comments violating these policies, of which 15,700 were posted before Meta's policy changes and 39,900 were posted after. This represents an overall more than doubling of harassment.<sup>34</sup>

Comments in this category include calls for politicians to commit suicide. Many target politicians, and particularly women, with sexualized commentary as well as gendered slurs such as "whores" or "sluts".

Comments attempting to "doxx" politicians by revealing sensitive information such as home addresses or phone numbers, or calling on others to do so, were relatively rare.

"Hey Marjorie are you going to be sucking on Trump's bloody ear today so you can get an orgasm you sick type of clown you"

– Targeting Marjorie Taylor Greene

"Wipe your chin, Nancy, you got some of his juice on your face, still."

– Targeting Nancy Mace

"YOU STFU YOU DONT SPEAK FOR TEXANS GET YOUR FUCKING SLIMMY SKANK ASS BACK TO ST LOUIS YOU SOROS WHORE"

– Targeting Jasmine Crockett

"Current Address: \*\*\*\*\*  
Past Addresses: \*\*\*\*\* +16  
more Phone Number: (\*\*\*) \*\*\*-  
\*\*\*\* +4 phones  
Email Address: \*\*\*\*\*"

– Targeting Darin LaHood

"Robert Garcia – If you're so concerned about this country, you can best serve it by committing suicide."

– Targeting Robert Garcia

"I'm praying you get cancer."

– Targeting Mary Miller

"After seeing you on the news attacking Elon. It came to mind you looked like a rat. So I am just going to call you little rat boy"

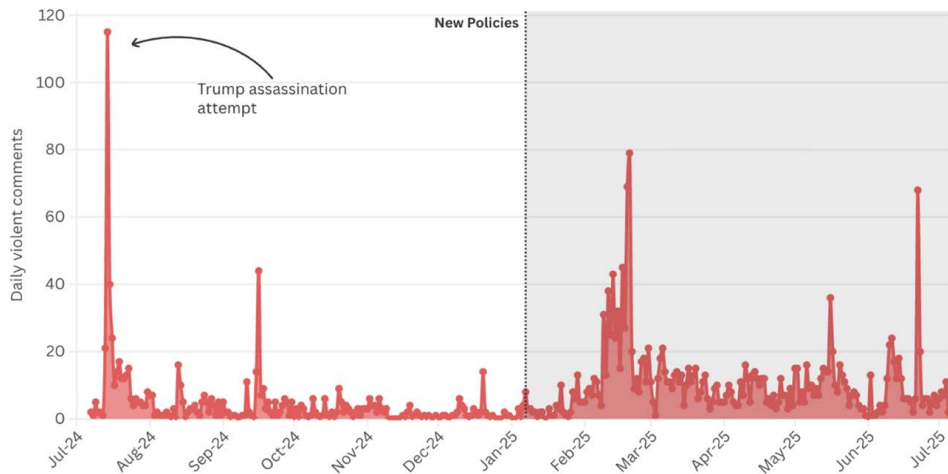
– Targeting Robert Garcia

"Stfu you traitor slut"

– Targeting Marjorie Taylor Greene

## 5. THREATS AGAINST TRUMP SURGED

Violence and incitement in comments targeting Trump doubled



We set out to analyze comments targeting Members of Congress, but quickly found many comments that contain threats against President Donald J. Trump.

Analysis of these comments found that threats against Trump were twice as high in the six months after Meta's policy changes than in the six months prior.

This analysis was designed to identify comments targeting Trump that also violate Meta's policies on violence and incitement.<sup>35</sup>

Our analysis identified an estimated 2,700 comments matching these criteria, of which 800 were posted before Meta's policy changes and 1,900 were posted after. This represents a doubling in threats against Trump.<sup>36</sup>

Trump has been the target of multiple assassination attempts.<sup>37</sup> The most notable occurred in July 2024 at a rally in Pennsylvania, and violent comments towards Trump spiked following the attack.<sup>38</sup>

Comments in this category included many explicit calls to kill Trump, references to the Pennsylvania assassination attempt that incite violence, and calls to target Trump at the White House or his residence in Mar-a-Lago.

"a bullet and a body bag trump is a walking dead man period"

"Damn, just two more inches to the right... #soclosebutsofar"

"This would be a good time for the Senate to be stormed in an active shooter. Find Donald Trump."

"4 inches to the left and the shooter would have Made America Great Again."

"YES , ASSASINATE TRUMP AND MTS LAG"

"Trump deserves a bullet through his head."

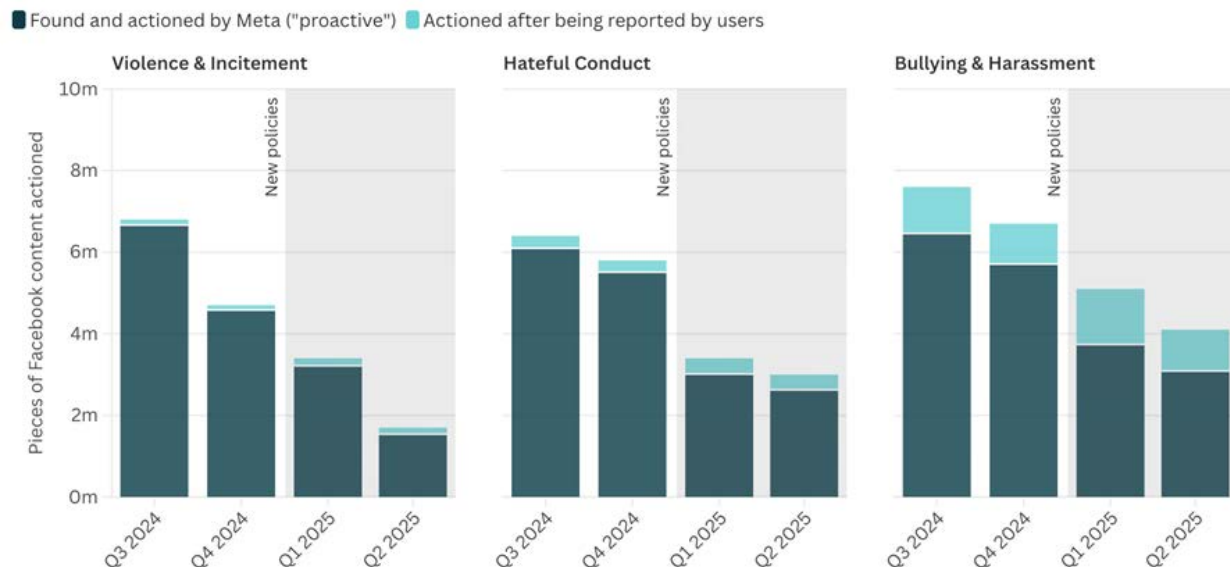
"Trump, Musk and every elected GOP member needs to be publicly executed"

"Donald trump must die to free America."

"THERE IS ALWAYS ANOTHER SOLUTION TO GET RID OFF THE MAGA TRAITOR LAWMAKERS AND THEIR ORANGE POS!!! TAKE YOUR GUNS AND MAKE THEIR LIVES A LIVING HELL!! GO TO THEIR HOMES AND AND DO WHAT TO DO!! SAY "HELLO" TO THEIR FAMILIES AND SHOW UP WERE EVER THEY ARE!!!! AND DONT FORGET YOUR GUNS!!!"

## 6. META CUT ENFORCEMENT

Meta halved enforcement on Facebook in the period that abuse surged



We find that abuse surged in the wake of Meta's policy changes, raising the question of which policy changes may have contributed to a rise in the volume of abuse.

One key policy change was that Meta would cut back on its use of "automated systems" to enforce its policies, something Meta calls "proactive enforcement", instead relying on users to report content.<sup>39</sup>

Meta's own transparency reports show this resulted in a sharp drop in Facebook's enforcement action against abusive content in the period that abuse surged.<sup>40</sup>

In the six months before Meta's policy changes, Facebook took proactive action on 35 million pieces of abusive content under its policies on violence and incitement, hateful conduct, and bullying and harassment.<sup>41</sup>

In the six months that followed the changes, Facebook took proactive action against just 17.2 million pieces of content, equivalent to a halving of Meta's proactive enforcement against abusive content.<sup>42</sup>

Enforcement in response to user reports increased slightly during the same period, with Facebook acting on 3 million user reports of abuse in the six months before the changes and 3.5 million after.<sup>43</sup>

This suggests that Meta's cuts to proactive enforcement may have contributed to the surge in abuse we identify in our analysis of comments directed at politicians.

# APPENDIX:

# METHODOLOGY

---

This appendix sets out our methodology in more detail, including how we selected politicians' accounts for study, how post and comment data were collected, and how AI systems were used to identify comments that violated Meta's platform policies.

## HOW WE SELECTED POLITICIANS TO STUDY

Our study analyzes comments on Facebook posts from the 50 Republicans and the 50 Democrats with the most followers in the US House of Representatives.

To arrive at this list of politicians, researchers first reviewed the list of members of the House of Representatives published on August 15, 2025.<sup>44</sup> Members who were not serving for the entire period of study spanning the six months before and after Meta's policy changes of January 7, 2025 were eliminated from our analysis, including:

- ▶ Members who were not in office before the election on November 5, 2024
- ▶ Members who stepped down during the period of study
- ▶ Members who passed away in the period of study

This resulted in a list of 370 members of the House of Representatives that were eligible for our study and matched to Facebook pages they operate in the next step.

## HOW WE SELECTED FACEBOOK PAGES TO STUDY

After compiling a list of members eligible for further study, researchers needed to identify the 50 Republicans and 50 Democrats with the most followers on Facebook.

They did so by compiling a list of Facebook pages operated by eligible members using automated Google site search queries and Facebook's own search tools. Researchers recorded each page's related member, its URL and number of followers. Where members operate multiple pages, all of them were collected.

Members were then ranked by the number of followers on their largest public Facebook page. The 50 Republicans and 50 Democrats with the most followers on their largest pages were taken forward for further data collection and analysis.

The resulting list of members is available in Appendix: List of Members Studied.

## HOW WE COLLECTED OUR COMMENT DATASET

Using our list of the 50 Republican and 50 Democrat members with the most Facebook followers, we collected comments on posts from all their pages in the period studied.

This list of 100 members operates a total of 187 public Facebook pages. For each page, researchers collected all posts made in the six months before and six months after Meta’s policy changes, a period spanning July 7, 2024 to July 6, 2025 inclusive. This resulted in a dataset of 63,994 posts made in the period of study.

For each post in our dataset, we then collected all first-level comments made in the same period of study spanning July 7, 2024 to July 6, 2025 inclusive. Second-level comments that respond to first-level comments were not collected in order to focus our study on comments that are most likely to be directly responding to politicians’ posts. This resulted in a dataset of 7,868,763 comments for analysis.

Data collection was carried out between October 15 and November 25, 2025. It was carried out using a data collection tool for Facebook posts and comments made by Bright Data, the leading web data infrastructure company, and was supported by the Bright Initiative by Bright Data.

Bright Data is trusted by over 20,000 organizations to ethically access and collect public web data at scale. Supporting data for good projects across the globe through the Bright Initiative by Bright Data, nonprofits, researchers, and public institutions are empowered with state of the art tools. 750+ global mission-driven partners have accessed the platform to expand the reach, efficiency, and effectiveness of their work across vital areas, in AI, education, climate action, human rights, and public health.

## **HOW WE CATEGORIZED COMMENTS VIOLATING META’S POLICIES ON ABUSE**

Facebook comments in our analysis were compared against Meta’s policies on abuse, which we define as comprising these three parts of Meta’s Community Standards:

1. Violence and incitement
2. Hateful conduct
3. Bullying and harassment

Comments were placed in one of the above categories if they target the politician whose post they appear on and violate Meta’s relevant policy. Throughout the report we collectively refer to any comments meeting these criteria as “abusive” comments.

Where comments violate more than one of the above policies, they were placed in just the highest-ranking category to simplify our analysis. For example, a comment violating Meta’s policies on violence and hate would be placed only in the violence category.

## **HOW WE ANALYZED COMMENTS USING AI**

Using the above definitions, we developed an AI system to identify and categorize abusive Facebook comments that target politicians and violate Meta’s policies.

This system works by prompting OpenAI’s large language models (LLMs) to assess Facebook comments in our dataset against Meta’s policies, including only those comments that target the politicians whose posts they are commenting on.

The system uses low-cost models to make initial assessments before graduating to high-cost models that are more accurate. This process was repeated for each relevant policy area: violence and incitement, hateful conduct, and bullying and harassment.

Our approach for analyzing violence and incitement involved an extra stage aimed at capturing comments directed at President Donald J. Trump as well as the 100 Members of Congress used as the basis for our core analysis.

The table below outlines the AI system used for our analysis, including how it differed for our analysis of violence and incitement.

| Violence and Incitement                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Hateful Conduct/Bullying and Harassment                                                                                                                                                                                                                                                                                                                                                                                                     |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Stage 1</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| OpenAI's GPT-4o-mini model identifies comments most relevant to each policy area. We used reverse engineered chain-of-thought prompting (RECoT) developed by Coan et al., which teaches a low-cost model "reasoning" skills based on relevant examples. <sup>45</sup> This process enabled classification of comments at a substantially lower cost than running a high-performing model across the full dataset.                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Stage 2</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Comments identified in the previous step were reassessed using OpenAI's more capable GPT-5.1 model with a stricter evaluation to determine whether comments violate Meta's policies. The model was not instructed to identify only comments that target the relevant politician at this stage.                                                                                                                                                                                                                     | Comments identified in the previous step were reassessed using OpenAI's more capable GPT-5.1 model with a stricter evaluation to determine whether comments violate Meta's policies. The model was also instructed to identify only comments that target the relevant politician, using additional contextual inputs including the relevant politician's name, party affiliation, post text, ethnicity, religion, and gender. <sup>46</sup> |
| <b>Stage 3</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Comments identified as violating Meta's violence and incitement policy in the previous stage were reassessed using a OpenAI's GPT-5.1 to determine their target. This enabled researchers to identify posts targeting President Donald J. Trump as well as those targeting the politician whose post contained the comment. The model was given additional context including the relevant politician's name, party affiliation, post text, ethnicity, religion, and gender to assist in assessments. <sup>47</sup> | <i>Not applicable.</i>                                                                                                                                                                                                                                                                                                                                                                                                                      |
| <b>Output</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Comments that target politicians and violate Meta's relevant policy.                                                                                                                                                                                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                             |

## HOW WE EVALUATED THE PERFORMANCE OF THE AI SYSTEM

We assessed the AI system outlined above for its performance in identifying Facebook comments that target politicians with abuse violating Meta’s policies.

This assessment takes a sample of comments relevant to each policy area and compares the AI system’s categorization of them against those made by researchers trained to apply Meta’s policies to comments targeting politicians. Where researchers disagreed on an assessment, they deliberated in a group to reach consensus.

Samples were extracted using stratified random sampling, with over-sampling of comments flagged by a preliminary AI assessment as more likely to violate policies. This approach ensured that policy violations were well represented in small samples.

Across all three policies the AI system achieved an average F1 score of 79%, an average precision score of 79%, and an average recall score of 80%, defined as follows:

- ▶ F1 score: the harmonic mean of Precision and Recall.
- ▶ Precision: the percentage of posts labelled as belonging to a category by the model that were actually correct.
- ▶ Recall: the percentage of posts for a category that were correctly identified by the model as belonging to that category.

Accurately identifying and categorizing content from social media is a complex task, and while our AI system reaches a high level of accuracy in this task, it was not perfectly accurate. This is typical when using AI to categorize large numbers of posts.<sup>48</sup>

The accuracy of the AI system differed slightly across each policy area studied, with scores summarized in the table below. Scores are based on a weighted analysis of each sample, adjusted to reflect the underlying population distribution. The table also states the size of the sample used to assess performance in each policy area.

| Category                                              | Precision | Recall | F1 Score | Sample |
|-------------------------------------------------------|-----------|--------|----------|--------|
| <i>Policy Areas</i>                                   |           |        |          |        |
| Violence and Incitement                               | 75%       | 73%    | 74%      | 2,339  |
| Hateful Conduct                                       | 80%       | 90%    | 85%      | 2,108  |
| Bullying and Harassment                               | 82%       | 78%    | 80%      | 2,040  |
| Overall Macro Average                                 | 79%       | 80%    | 79%      |        |
| <i>Trump Analysis (not included in macro average)</i> |           |        |          |        |
| Violence and Incitement targeting Trump               | 84%       | 65%    | 73%      | 2,339  |

## LIMITATIONS OF OUR STUDY

The methodology outlined in this appendix introduces limitations to our study, which are outlined below.

### *Our dataset excludes comments on posts made prior to our study period*

We collected only comments on posts made within the study period of July 7, 2024 to July 6, 2025 inclusive. This means the dataset does not include comments made within the study period but on posts that were made before the study period. Analysis shows that most comments on a post appear shortly after it is posted, indicating that this does not have a substantive impact on our findings.

### *Random errors in data collection*

Data collection was imperfect, with cases where posts or comments failed to collect due to processing errors. Our analysis indicates that the errors are randomized, with no relationship to particular time periods, Facebook pages or types of comment. As such, this issue should not have a substantive impact on our findings.

To minimize the effect of these randomized errors, researchers repeated post collection twice. Where data collection on posts yielded fewer comments than we would expect, we also repeated comment collection three times.

# APPENDIX:

## LIST OF MEMBERS STUDIED

This appendix lists the Members of Congress selected for study. Appendix: Methodology explains in detail how these politicians were selected.

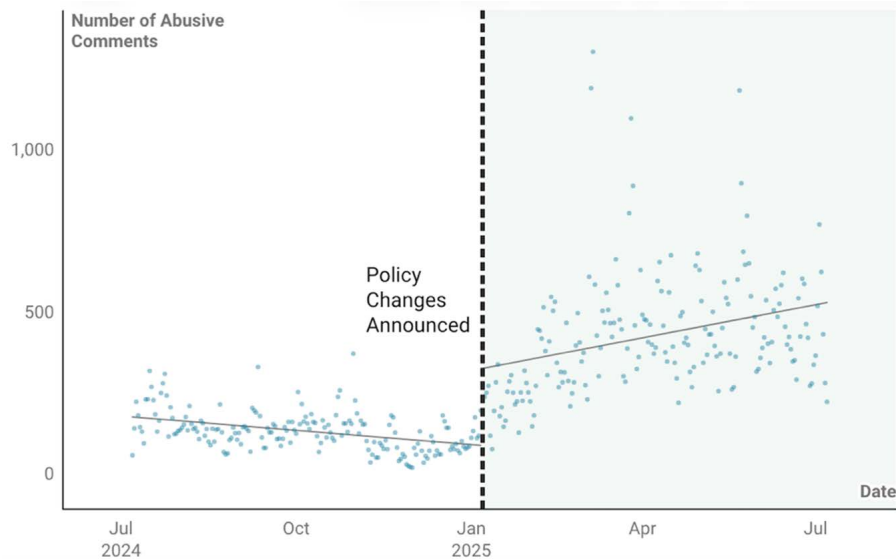
| Politician Name        | Party |
|------------------------|-------|
| Aderholt, Robert B.    | R     |
| Allen, Rick W.         | R     |
| Arrington, Jodey C.    | R     |
| Babin, Brian           | R     |
| Balint, Becca          | D     |
| Beatty, Joyce          | D     |
| Beyer, Donald S., Jr.  | D     |
| Biggs, Andy            | R     |
| Bilirakis, Gus M.      | R     |
| Boebert, Lauren        | R     |
| Bost, Mike             | R     |
| Brown, Shontel M.      | D     |
| Buchanan, Vern         | R     |
| Burchett, Tim          | R     |
| Cammack, Kat           | R     |
| Carter, Earl L., Buddy | R     |
| Castro, Joaquin        | D     |
| Clark, Katherine M.    | D     |
| Clyburn, James E.      | D     |
| Cohen, Steve           | D     |
| Comer, James           | R     |
| Crenshaw, Dan          | R     |
| Crockett, Jasmine      | D     |
| Crow, Jason            | D     |
| DesJarlais, Scott      | R     |
| Dingell, Debbie        | D     |
| Doggett, Lloyd         | D     |
| Donalds, Byron         | R     |
| Fallon, Pat            | R     |
| Garamendi, John        | D     |
| Garcia, Robert         | D     |
| Gimenez, Carlos A.     | R     |
| Gomez, Jimmy           | D     |

| Politician Name         | Party |
|-------------------------|-------|
| Gooden, Lance           | R     |
| Gosar, Paul A.          | R     |
| Graves, Sam             | R     |
| Green, Al               | D     |
| Greene, Marjorie Taylor | R     |
| Higgins, Clay           | R     |
| Himes, James A.         | D     |
| Hoyer, Steny H.         | D     |
| Huizenga, Bill          | R     |
| Hunt, Wesley            | R     |
| Issa, Darrell           | R     |
| Jackson, Ronny          | R     |
| James, John             | R     |
| Jayapal, Pramila        | D     |
| Jeffries, Hakeem S.     | D     |
| Johnson, Mike           | R     |
| Jordan, Jim             | R     |
| Khanna, Ro              | D     |
| Krishnamoorthi, Raja    | D     |
| LaHood, Darin           | R     |
| Larson, John B.         | D     |
| Lieu, Ted               | D     |
| Lucas, Frank D.         | R     |
| Luna, Anna Paulina      | R     |
| Mace, Nancy             | R     |
| Malliotakis, Nicole     | R     |
| Massie, Thomas          | R     |
| Mast, Brian J.          | R     |
| McClain, Lisa C.        | R     |
| McClintock, Tom         | R     |
| Menendez, Robert        | D     |
| Miller, Mary E.         | R     |
| Moulton, Seth           | D     |

| Politician Name           | Party |
|---------------------------|-------|
| Norman, Ralph             | R     |
| Norton, Eleanor Holmes    | D     |
| Ocasio-Cortez, Alexandria | D     |
| Omar, Ilhan               | D     |
| Pelosi, Nancy             | D     |
| Pocan, Mark               | D     |
| Pressley, Ayanna          | D     |
| Ramirez, Delia C.         | D     |
| Raskin, Jamie             | D     |
| Roy, Chip                 | R     |
| Salazar, Maria Elvira     | R     |
| Scalise, Steve            | R     |
| Schakowsky, Jan D.        | D     |
| Schweikert, David         | R     |
| Scott, Austin             | R     |
| Sewell, Terri A.          | D     |
| Sherman, Brad             | D     |
| Sherrill, Mikie           | D     |
| Smith, Jason              | R     |
| Sorensen, Eric            | D     |
| Stansbury, Melanie A.     | D     |
| Stanton, Greg             | D     |
| Stefanik, Elise M.        | R     |
| Steube, W. Gregory        | R     |
| Swalwell, Eric            | D     |
| Takano, Mark              | D     |
| Thanedar, Shri            | D     |
| Thompson, Bennie G.       | D     |
| Tlaib, Rashida            | D     |
| Velázquez, Nydia M.       | D     |
| Wasserman Schultz, Debbie | D     |
| Waters, Maxine            | D     |
| Wilson, Joe               | R     |

# APPENDIX: STATISTICAL SIGNIFICANCE

## Interrupted time series analysis shows a step-change in abuse



This appendix uses an interrupted time series analysis to find that the surge in abuse directed at politicians after Meta’s policy changes is statistically significant.

The interrupted time series model estimates three key parameters: the pre-existing trend in daily abusive comments, an immediate level shift at the point of Meta’s policy changes, and a change in the ongoing trend after the changes. Overall, the model fit the data well, explaining approximately 58% of the variance in daily abusive comments ( $R^2 = 0.58$ ,  $F(3, 362) = 78.76$ ,  $p < 0.001$ ).

At the start of the study period, the estimated daily volume of abusive comments was around 177 comments per day, with a statistically significant downward trend of approximately 0.48 comments per day ( $\beta = -0.48$ , 95% CI  $[-0.709, -0.250]$ ,  $p < 0.001$ ). This suggests that abusive comments were gradually declining before Meta’s policy changes.

Following Meta’s policy changes, there was a statistically significant increase in the volume of abusive comments. The model estimates an increase of approximately 236 additional abusive comments per day at the point the changes were announced ( $\beta = 236.12$ , 95% CI  $[152.00, 320.24]$ ,  $p < 0.001$ ), representing a more than doubling of average daily abuse.

Following this abrupt level change, the volume of abusive comments continued to grow at a statistically significant rate. The period after the policy changes shows an upward trend of approximately 1.6 additional comments per day ( $\beta = 1.601$ , 95% CI  $[0.862, 2.341]$ ,  $p < 0.001$ ), representing a real and sustained rise in abuse that has persisted throughout the observation period.

## HOW WE APPLIED INTERRUPTED TIME SERIES ANALYSIS TO COMMENT DATA

The unit used for this analysis is the daily count of abusive comments across all 100 politicians' accounts. These counts are estimates based on our AI analysis, and "abusive comments" are those violating Meta's policies on violence and incitement, hateful conduct, and bullying and harassment.

The interruption point was set at January 7, 2025, the date of Meta's policy changes.<sup>49</sup> While Meta confirmed that changes to its Community Standards took effect that same day, it did not specify implementation dates for the other changes.<sup>50</sup>

The model is structured around three variables representing key parameters of interest:

1. **A pre-intervention time variable:** this captures the underlying trend in daily abusive comments in the six months prior to the intervention (i.e. the policy changes), representing the counterfactual trajectory against which any change after that date is measured.
2. **A binary indicator variable:** coded 0 for all days before the intervention and 1 for all days after. This captures any immediate and sustained shift in the level of abusive comments at the point of the intervention.
3. **A post-intervention time variable:** coded 0 before the intervention and counting upward from 1 after it, estimating whether the rate of change in daily abusive comments shifted following the intervention.

Together, these variables allow the model to estimate both how abusive comment volumes were trending before the policy changes were announced, and whether they were associated with an abrupt change in level, a change in ongoing rate of growth, or both. The model also estimates an intercept, representing the baseline level of daily abusive comments at the start of the observation period.

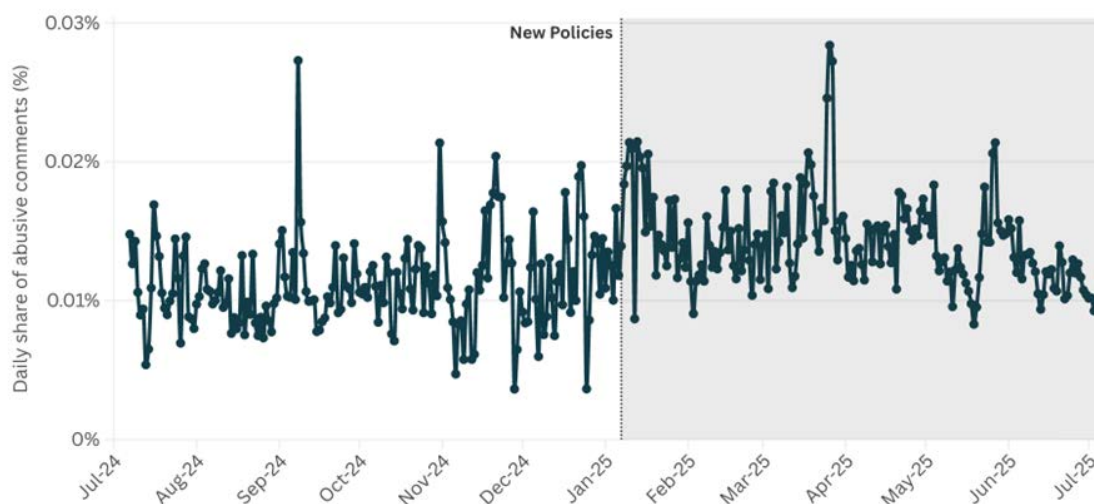
The segmented regression model was estimated using ordinary least squares (OLS) regression using the Newey–West HAC standard errors model, which adjusts for two known limitations of ITS analysis: autocorrelation and heteroscedasticity.<sup>51</sup>

Autocorrelation is where observations close together tend to be correlated and heteroscedasticity refers to unequal variance across the observation period. Both can cause standard errors to be underestimated, making findings appear more statistically significant than they are.

# APPENDIX:

## SHARE OF ABUSIVE COMMENTS

Abusive comments as a share of all comments increased from 1.09% to 1.38%



This report finds a surge in the overall volume of abusive comments targeting politicians on Facebook since Meta’s policy changes on January 7, 2025.<sup>52</sup> To further understand the nature of this change, it is important to put it into the context of an overall increase in comments on politicians’ posts over the same period.

Our data shows that the overall volume of comments on politicians’ posts rose to 5.6 million in the six months after Meta’s policy changes compared to 2.2 million comments in the six months before. At the same time, there was also a rise in the share of those comments that were abusive, from 1.09% in the six months before compared to 1.38% in the six months after. The daily share of abusive comments is shown above.

The policy areas studied differed in terms of the amount they changed as a share of all comments. Comments containing violence and incitement increased from 0.08% before the changes to 0.14% after. Hateful comments increased from 0.31% before the changes compared to 0.53% after. Bullying and harassing comments took a different path, remaining essentially stable but increasing slightly from 0.7% to 0.71%.

This suggests that our finding of a surge in the volume of abuse could be a combination of two factors: an increase in overall comments on politicians’ posts, and an increase in the share of those comments that were abusive in nature. This implies that not only are politicians experiencing a surge in abuse driven by increased engagement, but that a higher share of that engagement has become abusive.

Notably, these findings are consistent with two key elements of Meta's policy changes. First, Meta's stated it would be reversing a previous decision to reduce visibility of political content on people's feeds, meaning more political content would be recommended based on user signals such as post likes.<sup>53</sup> This change may have resulted in higher visibility of politicians' posts and greater overall engagement.

Second, Meta announced a roll back the use of automated systems to detect and act on violating content – also known as “proactive” enforcement.<sup>54</sup> Such a reduction in policy enforcement could explain an increase in the overall share of abusive content.

Taken together, increasing the visibility of politicians' posts while reducing proactive enforcement could be expected to generate a surge in abuse on politicians' posts.

### **INCREASE IN OVERALL SHARE OF ABUSIVE COMMENTS IS STATISTICALLY SIGNIFICANT**

As with the volume of abusive comments, an interrupted time series analysis was conducted on the daily proportion of comments estimated to be abusive, using the same model structure and interruption point as described above. The model, which was run using Newey–West HAC standard errors to adjust for autocorrelation and heteroscedasticity, explained approximately 23% of the variance in the daily proportion of abusive comments ( $R^2 = 0.227$ ,  $F(3, 362) = 17.47$ ,  $p < 0.001$ ).

Following Meta's policy changes, there was a statistically significant increase in the proportion of comments estimated to be abusive. The model estimates an increase of approximately 0.4% at the point the changes were announced ( $\beta = 0.0040$ , 95% CI [0.002, 0.006],  $p = <0.001$ ). Prior to the policy changes, the proportion of abusive comments was broadly stable, with no statistically significant pre-existing trend ( $\beta = 9.208e-06$ , 95% CI [-1.21e-06, 1.96e-05],  $p = 0.083$ ). In the period after the policy changes, the proportion showed a small but statistically significant downward trajectory ( $\beta = -3.01e-05$ , 95% CI [-4.65e-05, -1.37e-05],  $p = <0.001$ ).

# REFERENCES

- <sup>1</sup>“Man charged in White House correspondents’ dinner attack pleads not guilty to assassination attempt”, Michael Kunzleman, PBS, 11 May 2026, <https://www.pbs.org/newshour/nation/man-charged-in-white-house-correspondents-dinner-at-tack-pleads-not-guilty-to-assassination-attempt>
- <sup>2</sup>“Man charged in White House correspondents’ dinner attack pleads not guilty to assassination attempt”, Michael Kunzleman, PBS, 11 May 2026, <https://www.pbs.org/newshour/nation/man-charged-in-white-house-correspondents-dinner-at-tack-pleads-not-guilty-to-assassination-attempt>
- <sup>3</sup>“Man who set fire to Pennsylvania governor’s mansion sentenced”, Ana Faguy, BBC, 14 October 2025, <https://www.bbc.co.uk/news/articles/cdegnne8r3go>
- <sup>4</sup>“A ‘formidable public servant.’ Who was Melissa Hortman, the Minnesota state representative assassinated in her home?”, Chris Boyette, CNN, 15 June 2025, <https://edition.cnn.com/2025/06/14/us/melissa-hortman-minnesota-assassination>
- <sup>5</sup>“Manhunt for shooter continues after Charlie Kirk killed in ‘political assassination’”, Meredith Deliso, ABC, <https://abcnews.com/US/charlie-kirk-shot-event-utah-university-jd-vance/story?id=125451514>
- <sup>6</sup>“Arizona Senator cancels town halls over recent security threats”, FOX, 20 September 2025, <https://www.fox10phoenix.com/video/1711207>
- <sup>7</sup>“U.S. Rep. Hageman of Wyoming cancels in-person town hall meetings”, MTN News, KTVQ, 26 March 2025, <http://ktvq.com/news/u-s-rep-hageman-of-wyoming-cancels-in-person-town-hall-meetings>
- <sup>8</sup>“Harassment Is Driving Local Election Officials in Swing States Out of Their Jobs”, Nuha Dolby, NOTUS, 16 October 2024, <https://www.notus.org/2024-election/election-clerk-turnover>
- <sup>9</sup>“Democrats voice fears of violence over Charlie Kirk vote”, Andrew Solender, Axios, 18 September 2025, <https://www.axios.com/2025/09/18/charlie-kirk-resolution-vote-democrats-violence>
- <sup>10</sup>“Mark Zuckerberg Stands for Voice and Free Expression”, Newsroom, Meta, 17 October 2019, <https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/>
- <sup>11</sup>“More Speech and Fewer Mistakes”, Newsroom, Joel Kaplan, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>
- <sup>12</sup>AI analysis of comments in our dataset identified an estimated total of 9,369 comments as violating Meta’s policies on violence and incitement, including 1,760 posted in the six months before Meta’s announcement of policy changes and 7,609 that were posted in the six months that followed. Our AI analysis is not perfectly accurate making these figures estimates. More detail is available in Appendix: Methodology.
- <sup>13</sup>AI analysis of comments in our dataset identified an estimated total of 36,897 comments as violating Meta’s policies on hateful conduct, including 6,875 posted in the six months before Meta’s announcement of policy changes and 30,022 that were posted in the six months that followed. Our AI analysis is not perfectly accurate making these figures estimates. More detail is available in Appendix: Methodology.
- <sup>14</sup>AI analysis of comments in our dataset identified an estimated total of 55,645 comments as violating Meta’s policies on bullying and harassment, including 15,730 posted in the six months before Meta’s announcement of policy changes and 39,915 that were posted in the six months that followed. Our AI analysis is not perfectly accurate making these figures estimates. More detail is available in Appendix: Methodology.
- <sup>15</sup>AI analysis of comments in our dataset identified an estimated total of 2,683 comments targeting President Donald J. Trump by violating Meta’s policies on violence and incitement, including 772 posted in the six months before Meta’s announcement of policy changes and 1,911 that were posted in the six months that followed. This means that 2.5 times as many abusive comments were posted in the period after the policy changes than before. Our AI analysis is not perfectly accurate making these figures estimates. More detail is available in Appendix: Methodology.
- <sup>16</sup>Bullying and Harassment, Community Standards Enforcement Report, Transparency Center, Meta, accessed 2 June 2026, <http://transparency.meta.com/reports/community-standards-enforcement/bullying-and-harassment/>
- Hateful Conduct, Community Standards Enforcement Report, Transparency Center, Meta, accessed 2 June 2026, <https://transparency.meta.com/reports/community-standards-enforcement/hateful-conduct/>
- Violence and Incitement, Community Standards Enforcement Report, Transparency Center, Meta, accessed 2 June 2026, <https://transparency.meta.com/reports/community-standards-enforcement/violence-incitement/>
- <sup>17</sup>“More Speech and Fewer Mistakes”, Meta, January 7, 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>
- <sup>18</sup>“More Transparency and Less Spin”, Center for Countering Digital Hate, February 24, 2025, <https://counterhate.com/research/more-transparency-and-less-spin/>
- <sup>19</sup>“More Speech and Fewer Mistakes”, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>
- <sup>20</sup>“More Transparency and Less Spin”, Center for Countering Digital Hate, February 24, 2025, <https://counterhate.com/research/more-transparency-and-less-spin/>
- <sup>21</sup>See Section 6 of this report for more detail.
- <sup>22</sup>AI analysis of comments in our dataset identified a total of 101,911 comments as abusive, including 24,365 posted in the six months before Meta’s announcement of policy changes and 77,546 that were posted in the six months that followed. More detail is available in Section 4: “Abuse surged after policy changes”.
- <sup>23</sup>The data collected spans one year, covering the six months before and six months after Meta’s policy changes on January 7, 2025, i.e. from and including July 7, 2024 to July 6, 2025. Researchers collected all comments from within this time window that appeared on posts created within the time window. The six-month time-period before the changes was defined as from July 7, 2024 to January 6, 2025, and the six-month time-period after the changes was defined as from January 7, 2025, to July 6, 2025.
- <sup>24</sup>“Violence and incitement”, Meta, Community Standards, accessed February 5, 2026, <https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/>
- <sup>25</sup>“Hateful conduct”, Meta, Community Standards, accessed February 5, 2026, <https://transparency.meta.com/policies/community-standards/hateful-conduct/>
- <sup>26</sup>“Bullying and harassment”, Meta, Community Standards, accessed February 5, 2026, <https://transparency.meta.com/en-gb/policies/community-standards/bullying-harassment/>
- <sup>27</sup>More detail on this is available in Section: “Appendix: Methodology”
- <sup>28</sup>“Violence and incitement”, Meta, Community Standards, accessed February 5, 2026, <https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/>
- <sup>29</sup>“Hateful conduct”, Meta, Community Standards, accessed February 5, 2026, <https://transparency.meta.com/policies/community-standards/hateful-conduct/>
- <sup>30</sup>“Bullying and harassment”, Meta, Community Standards, accessed February 5, 2026, <https://transparency.meta.com/en-gb/policies/community-standards/bullying-harassment/>
- <sup>31</sup>AI analysis of comments in our dataset identified an estimated total of 101,911 comments as abusive, including 24,365 posted in the six months before Meta’s announcement of policy changes and 77,546 that were posted in the six months that followed. This means that 3.2 times as many abusive comments were posted in the period after the policy changes than before. Our AI analysis is not perfectly accurate, making these figures estimates. More detail is available in Appendix: Methodology.
- <sup>32</sup>AI analysis of comments in our dataset identified an estimated total of 9,369 comments as violating Meta’s policies on violence and incitement, including 1,760 posted in the six months before Meta’s announcement of policy changes and 7,609 that were posted in the six months that followed. This means that 4.3 times as many abusive comments were posted in the period after the policy changes than before. Our AI analysis is not perfectly accurate making these figures estimates. More detail is available in Appendix: Methodology.
- <sup>33</sup>“Violence and incitement”, Meta, accessed March 5, 2026, <https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/>
- <sup>34</sup>“Violence and incitement”, Meta, accessed March 5, 2026, <https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/>
- <sup>35</sup>AI analysis of comments in our dataset identified an estimated total of 9,369 comments as violating Meta’s policies on violence and incitement, including 1,760 posted in the six months before Meta’s announcement of policy changes and 7,609 that were posted in the six months that followed. This means that 4.3 times as many abusive comments were posted in the period after the policy changes than before. Our AI analysis is not perfectly accurate making these figures estimates. More detail is available in Appendix: Methodology.
- <sup>36</sup>“Hateful conduct”, Meta, accessed March 5, 2026, <https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/>
- <sup>37</sup>“And we know from history that dehumanizing people is the first step towards inciting violence. If you say immigrants are vermin, or all Muslims are terrorists — that makes others feel they can escalate and attack that group without consequences. So we don’t allow that. I take this incredibly seriously, and we work hard to get this off our platform.”
- <sup>38</sup>“Mark Zuckerberg Stands for Voice and Free Expression”, Meta, October 17, 2019, <https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/>

<sup>32</sup> AI analysis of comments in our dataset identified an estimated total of 36,897 comments as violating Meta's policies on hateful conduct, including 6,875 posted in the six months before Meta's announcement of policy changes and 30,022 that were posted in the six months that followed. This means that 4.4 times as many abusive comments were posted in the period after the policy changes than before. Our AI analysis is not perfectly accurate making these figures estimates. More detail is available in Appendix: Methodology.

<sup>33</sup> "Bullying and harassment", Meta, accessed March 5, 2026, <https://transparency.meta.com/policies/community-standards/bullying-harassment/>

<sup>34</sup> AI analysis of comments in our dataset identified an estimated total of 55,645 comments as violating Meta's policies on bullying and harassment, including 15,730 posted in the six months before Meta's announcement of policy changes and 39,915 that were posted in the six months that followed. This means that 2.5 times as many abusive comments were posted in the period after the policy changes than before. Our AI analysis is not perfectly accurate making these figures estimates. More detail is available in Appendix: Methodology.

<sup>35</sup> "Violence and incitement", Meta, accessed March 11, 2026, <https://transparency.meta.com/en-gb/policies/community-standards/violence-incitement/>

<sup>36</sup> AI analysis of comments in our dataset identified an estimated total of 2,683 comments targeting President Donald J. Trump by violating Meta's policies on violence and incitement, including 772 posted in the six months before Meta's announcement of policy changes and 1,911 that were posted in the six months that followed. This means that 2.5 times as many abusive comments were posted in the period after the policy changes than before. Our AI analysis is not perfectly accurate making these figures estimates. More detail is available in Appendix: Methodology.

<sup>37</sup> "Trump target of assassination attempt; says he was shot in ear at rally", NBC, Dasha Burns, Jake Traylor, Megan Lebowitz, Chloe Atkins, Tom Winter and Jonathan Dienst, July 13, 2024, <https://www.nbcnews.com/politics/donald-trump/secret-service-rushes-trump-stage-shots-fired-pennsylvania-rally-rcna161735>

"British man pleads guilty to plan to shoot Trump at Las Vegas rally", The Guardian, September 13, 2016, <https://www.theguardian.com/us-news/2016/sep/13/michael-steven-sandford-guilty-trump-shooting-attempt-las-vegas>

<sup>38</sup> "Trump target of assassination attempt; says he was shot in ear at rally", NBC, Dasha Burns, Jake Traylor, Megan Lebowitz, Chloe Atkins, Tom Winter and Jonathan Dienst, July 13, 2024, <https://www.nbcnews.com/politics/donald-trump/secret-service-rushes-trump-stage-shots-fired-pennsylvania-rally-rcna161735>

<sup>39</sup> "More Speech and Fewer Mistakes", Meta, January 7, 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

<sup>40</sup> "Community Standards Enforcement Report: H2 2025 Report", Meta, accessed February 11, 2026, <https://transparency.meta.com/reports/community-standards-enforcement/>

<sup>41</sup> This is calculated based on Meta's figures from Q3 and Q4 2024 for the total number of pieces of content that Facebook took action on across its policies on violence and incitement, hateful conduct and bullying and harassment, multiplied by the "proactive rate", meaning the share of content that was found and acted on by Facebook before being reported by users. The precise figure is 35,024,100.

"Community Standards Enforcement Report: H2 2025 Report", Meta, accessed February 11, 2026,

<https://transparency.meta.com/reports/community-standards-enforcement/>

<sup>42</sup> This is calculated based on Meta's figures from Q1 and Q2 2025 for the total number of pieces of content that Facebook took action on across its policies on violence and incitement, hateful conduct and bullying and harassment, multiplied by the "proactive rate", meaning the share of content that was found and acted on by Facebook before being reported by users. The precise figure is 17,219,600.

"Community Standards Enforcement Report: H2 2025 Report", Meta, accessed February 11, 2026,

<https://transparency.meta.com/reports/community-standards-enforcement/>

<sup>43</sup> This is calculated based on comparing Meta's figures from Q3-Q4 2024 and Q1-Q2 2025. It is based on Meta's figures for the total number of pieces of content that Facebook took action on across its policies on violence and incitement, hateful conduct and bullying and harassment, minus the number that were enforced proactively according to Meta's figures for the "proactive rate", meaning the share of content that was found and acted on by Facebook before being reported by users. The precise figures are 2,975,900 in Q3-Q4 2024 and 3,480,400 in Q1-Q2 2025.

"Community Standards Enforcement Report: H2 2025 Report", Meta, accessed February 11, 2026,

<https://transparency.meta.com/reports/community-standards-enforcement/>

<sup>44</sup> "119th Congress United States House of Representatives", Office of the Clerk of the United States House of Representatives, accessed August 15, 2025, [https://clerk.house.gov/member\\_info/TTD-119.pdf](https://clerk.house.gov/member_info/TTD-119.pdf)

<sup>45</sup> "Decoding Delay: Three Decades of Climate Change Opposition in the United States Congress", Travis G. Coan, Ranadheer Malla, Mirjam O. Nanko, William Kattrup, J. Timmons Roberts, John Cook, and Constantine Boussalis, accessed January 5, 2026, <https://github.com/project-c3ds/cards-2pO-paper>

<sup>46</sup> The following sources were used to collect information on the politicians' ethnicities, religions, party affiliation and gender:

"Equality PAC Marks the Start of the 119th Congress with a Record Number of LGBTQ Members", Equality PAC, January 3, 2025, <https://lgbtequalitypac.org/equality-pac-marks-the-start-of-the-119th-congress-with-a-record-number-of-lgbtq-members/>

"Black-American Members by Congress", United States House of Representatives, Office of the Historian, accessed October 25, 2025, <https://history.house.gov/Exhibitions-and-Publications/BAIC/Historical-Data/Black-American-Representatives-and-Senators-by-Congress/>

"Hispanic-American Representatives, Senators, Delegates, and Resident Commissioners by Congress", United States House of Representatives, Office of the Historian, accessed October 25, 2025, <https://history.house.gov/Exhibitions-and-Publications/HAIC/Historical-Data/Hispanic-American-Representatives-Senators-Delegates-and-Resident-Commissioners-by-Congress/>

"Asian and Pacific Islander American Representatives, Senators, Delegates, and Resident Commissioners by Congress", United States House of Representatives, Office of the Historian, accessed October 25, 2025, <https://history.house.gov/Exhibitions-and-Publications/APA/Historical-Data/By-Congress/>

"Jewish Members of the 119th Congress", Jewish Virtual Library, accessed October 25, 2025, <https://jewishvirtuallibrary.org/jewish-members-of-the-119th-congress>

"Women Elected Officials Database", Center for American Women and Politics, accessed October 25, 2025, <https://cawp.rutgers.edu/data/women-elected-officials-database>

"Foreign Born in the United States House of Representatives One Hundred Nineteenth Congress", March 2, 2026, [https://clerk.house.gov/documents/Foreign\\_Born.pdf](https://clerk.house.gov/documents/Foreign_Born.pdf)

<sup>47</sup> The following sources were used to collect information on the politicians' ethnicities, religions, party affiliation and gender:

"Equality PAC Marks the Start of the 119th Congress with a Record Number of LGBTQ Members", Equality PAC, January 3, 2025, <https://lgbtequalitypac.org/equality-pac-marks-the-start-of-the-119th-congress-with-a-record-number-of-lgbtq-members/>

"Black-American Members by Congress", United States House of Representatives, Office of the Historian, accessed October 25, 2025, <https://history.house.gov/Exhibitions-and-Publications/BAIC/Historical-Data/Black-American-Representatives-and-Senators-by-Congress/>

"Hispanic-American Representatives, Senators, Delegates, and Resident Commissioners by Congress", United States House of Representatives, Office of the Historian, accessed October 25, 2025, <https://history.house.gov/Exhibitions-and-Publications/HAIC/Historical-Data/Hispanic-American-Representatives-Senators-Delegates-and-Resident-Commissioners-by-Congress/>

"Asian and Pacific Islander American Representatives, Senators, Delegates, and Resident Commissioners by Congress", United States House of Representatives, Office of the Historian, accessed October 25, 2025, <https://history.house.gov/Exhibitions-and-Publications/APA/Historical-Data/By-Congress/>

"Jewish Members of the 119th Congress", Jewish Virtual Library, accessed October 25, 2025, <https://jewishvirtuallibrary.org/jewish-members-of-the-119th-congress>

"Women Elected Officials Database", Center for American Women and Politics, accessed October 25, 2025, <https://cawp.rutgers.edu/data/women-elected-officials-database>

"Foreign Born in the United States House of Representatives One Hundred Nineteenth Congress", March 2, 2026, [https://clerk.house.gov/documents/Foreign\\_Born.pdf](https://clerk.house.gov/documents/Foreign_Born.pdf)

<sup>48</sup> "Can GPT-4 detect subcategories of hatred?", R. U. Mustafa, N. Ashraf and N. Japkowicz, 2024 IEEE Digital Platforms and Societal Harms (DPSH), Washington, DC, USA, 2024, pp. 16, doi: 10.1109/DPSH60098.2024.10775211

<sup>49</sup> "More Speech and Fewer Mistakes", Meta, January 7, 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

<sup>50</sup> Jack Parrock: “You mentioned it a little bit there Joel, the changes are starting in the United States. When can we expect it in Europe? That’s what everyone wants to know here. You said 2026 just then – is that when we can expect it, from 2026 you’ll start rolling this out in the European Union as well?” Joel Kaplan, Chief Global Affairs Officer at Meta: “Yeah so I want to be clear, that was a reference to the Community Notes system that we’re replacing our third-party fact-checking program with. 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 US 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 US 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 US, and then we’ll expand from there.”

“Joel Kaplan on EU Regulation and Innovation”, Meta, 05:44, 7 February 2025, <https://about.fb.com/news/2025/02/joel-kaplan-on-eu-regulation-and-innovation/>

<sup>51</sup> Whitney K. Newey, Kenneth D. West, A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica*, 1987, <https://www.jstor.org/stable/1913610>

<sup>52</sup> “More Speech and Fewer Mistakes”, Meta, January 7, 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

<sup>53</sup> “More Speech and Fewer Mistakes”, Meta, January 7, 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

<sup>54</sup> “More Speech and Fewer Mistakes”, Meta, January 7, 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

