

MORE TRANSPARENCY AND LESS SPIN

Analyzing Meta's sweeping policy
changes and their impact on users.



Contents

Introduction	4
Executive Summary.....	5
Meta has questions to answer over sweeping changes announced on January 7th.....	7
1 Meta will halt “proactive” enforcement of some policies on harmful content.....	8
Meta is not clear which policies will no longer receive proactive enforcement.....	9
Meta could halt 97% of its enforcement in key policy areas such as hate speech.....	10
Table: Meta’s proactive enforcement from Q4 2023 to Q3 2024.....	11
Questions to answer on proactive enforcement.....	11
2 Meta will demote less content “that might violate our standards”	13
Questions to answer on content demotions.....	14
3 Meta is dropping policies on “immigration, gender identity and gender”	15
Questions to answer on policy changes.....	16
4 Meta will replace independent fact-checking with Community Notes.....	17
Community Notes perform poorly on polarizing, controversial topics.....	18
Questions to answer on fact-checking.....	19
5 Meta will demote less content about “elections, politics or social issues”	20
Questions to answer on civic content	20
6 Meta claims it will move trust and safety teams from California to Texas	21
Questions to answer on content moderation teams	21
References.....	22



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 non-profit company limited by guarantee.

Introduction

In 2018, Mark Zuckerberg's Facebook publicly accepted blame for allowing its platform to be used to "foment division and incite offline violence" that led to a genocide in Burma.¹ A sophisticated online propaganda campaign on Facebook conducted by the Burmese government, urging a national campaign of civilian slaughter, rape, and the mass displacement of entire communities, had been allowed to unfurl without consequences.



When such things happen, people solemnly promise each other that they will 'never forget'. For Mr. Zuckerberg, however, it appears six years was long enough. In January, Mr. Zuckerberg and his policy chief, Joel Kaplan, announced a new era for Meta, tearing up six years of slow, incremental change made in response not just to the Burmese genocide but a litany of sins that Meta had once acknowledged as its problem and had committed to solving. However, Mr. Zuckerberg and Mr. Kaplan failed to address what these changes would mean in practice in their breathless zeal to announce their change of heart.

To quantify the impact of Meta's newly announced policy platform, this new report by CCDH systematically reviews Meta's own data to piece together what we know, what we can safely predict, and we identify important unanswered questions that need urgent answers from Meta. Our analysis shows that the company's plans could affect 97% of its enforcement, which would result in a tidal wave of nearly 277 million pieces of hate speech and other harmful content flooding unabated onto the platform each year, threatening public safety, the integrity of our elections and democracy, as well as public health and the mental health of our children.

For the average user, this will mean that, when scrolling through Facebook, Instagram, or Threads, they are more likely to experience a chaotic blend of facts, opinion, and lies once Meta's proactive systems of detection of hate speech and misinformation are removed and its global fact-checking program is replaced by a copy of X's flawed Community Notes. Users are also far more likely to come across dehumanizing, hate-filled lies once prohibited by Meta, such as: "immigrants are grubby, filthy pieces of shit"; "Black people are more violent than Whites"; and "Jews are greedier than Christians". These aren't hypotheticals. These are specific examples cited in Meta's new internal guidance.²

Meta should have asked themselves all the questions raised in this report before announcing the changes. We urge legislators, regulators, civil society, journalists, citizens, and users to ask Meta these questions – whenever and however they can.

Imran Ahmed

CEO, Center for Countering Digital Hate

Executive Summary

Meta announced six key policy changes on January 7th

1. Halting “proactive” enforcement of some policies on harmful content
2. Demoting less content “that might violate our standards”
3. Dropping policies on “immigration, gender identity and gender”
4. Replacing independent fact-checking with Community Notes
5. Demoting less content about “elections, politics or social issues”
6. Moving trust and safety teams from California to Texas

Meta intends for these policy changes to be “expanded beyond the US”

- Meta’s Chief Global Affairs Officer, Joel Kaplan, has said that changes to fact-checking and enforcement will be “expanded beyond the US” in time.
- Kaplan also said that changes to Meta’s hate speech policies announced on January 7th “have been implemented worldwide immediately.”

1) Meta will halt “proactive” enforcement of some policies on harmful content

- Meta will halt proactive enforcement (including automatic detection) for some policies on harmful content, instead acting only in response to user reports.
- Meta’s announcement explicitly states proactive enforcement will continue for terrorism, child sexual exploitation, drugs, fraud and scams.
- Meta has not stated if proactive enforcement will continue for these policy areas used in Meta’s transparency reports, which we call “at risk” policy areas:
 - Bullying & Harassment
 - Dangerous Orgs: Organized Hate
 - Hate Speech
 - Suicide and Self-Injury
 - Violence And Incitement
 - Violence & Graphic Content
- Meta previously credited its “proactive detection technology” as a key factor in reducing the prevalence of hate speech and harmful content on its platforms.

Meta could halt 97% of its enforcement in key policy areas such as hate speech

- We analyzed Meta’s transparency reports to examine the potential impact of Meta halting proactive enforcement in policy areas such as hate speech.
- Last year, over 97% of Meta’s enforcement actions in “at risk” policy areas were “proactive”, with less than 3% made in response to user reports.
- Even accounting for Meta’s claims about mistakes in proactive enforcement, Meta correctly acted on 277 million pieces of content in “at risk” policy areas.

Meta must tell users which policies it will no longer proactively enforce, and how it will keep them safe if it stops acting on millions of pieces of harmful content.

2 Meta will demote less content “that might violate our standards”

- Meta will demote less content “that might violate our standards”, despite Mark Zuckerberg calling this as an “effective strategy” to reduce viral misinformation.
- In 2018, Mark Zuckerberg said Meta had evidence that the closer “borderline content” came to violating platform policies, the more engagement it gets.
- He said that demoting this content would reduce misinformation and “create a healthier, less polarized discourse where people feel safe participating.”

Meta must tell users why it is dropping an approach that they said reduces misinformation and polarization that is bad for users.

3 Meta is dropping policies on “immigration, gender identity and gender”

- Meta is dropping policies on “immigration, gender identity and gender”, with leaked internal guidance showing it will now permit statements such as:
 - “Trans people aren’t real. They’re mentally ill.”
 - “Immigrants are grubby, filthy pieces of shit.”
 - “Black people are more violent than Whites”
 - “Mexicans are lazier than Asians.”
 - “Jews are flat out greedier than Christians.”

Meta must tell users its assessment of the real-world consequences of this policy change and explain how it can continue to keep all of its users safe.

4 Meta will replace independent fact-checking with Community Notes

- Meta will replace independent fact-checking with Community Notes, and intends to roll out this new system worldwide after deploying it in the US.
- Community Notes can be useful, but a range of studies have found they perform poorly on divisive topics such as elections or viral misinformation.

Meta must tell users how it intends to address misinformation that undermines elections or poses a risk to the public given Community Notes’ known weaknesses.

5 Meta will demote less content about “elections, politics or social issues”

- Meta is reversing its 2021 decision to show users less “civic content”. At the time Meta claimed users did not want to see this content in their feeds.

Meta must tell users why it is making this change after saying users do not want it.

6 Meta claims it will move trust and safety teams from California to Texas

- Meta says it will move content moderation staff from California to Texas, but Meta has such teams in Texas and it is reported this change will affect few staff.

Meta must tell users if it plans to cut staffing on its US trust and safety teams.

Meta has questions to answer over sweeping changes announced on January 7th

On January 7th, Meta announced a raft of sweeping changes to its existing policies, including changes to fact-checking and its enforcement of policies on harmful content.

These changes were primarily announced in a blog titled “More Speech and Fewer Mistakes”, written by Joel Kaplan, Meta’s Chief Global Affairs Officer.³ This blog was accompanied by a five minute video of Mark Zuckerberg describing the changes, and a post, also by Zuckerberg, on the company’s platform Threads.⁴

Meta announced six key policy changes

For our analysis of Meta’s announcement and its potential impact, we have identified six key policy changes that are dealt with in subsequent sections of this report:

1. Halting “proactive” enforcement of some policies on harmful content
2. Demoting less content “that might violate our standards”
3. Dropping policies on “immigration, gender identity and gender”
4. Replacing independent fact-checking with Community Notes
5. Demoting less content about “elections, politics or social issues”
6. Moving trust and safety teams from California to Texas

Meta intends for these policy changes to “be expanded beyond the US”

Meta’s Chief Global Affairs Officer, Joel Kaplan, has said that changes to fact-checking and enforcement will be “expanded beyond the US” once they have been fully implemented in the US.⁵ Kaplan also said that changes to Meta’s hate speech policies (point three above) “have been implemented worldwide immediately.”⁶

Meta’s does not adequately explain these changes

Meta’s last earnings presentation reports that 3.35 billion people use its apps daily worldwide, each of whom helped Meta earn \$164 billion in revenue last year.⁷ In the US, polling indicates that 70% of adults use Facebook and 50% use Instagram.⁸

When these users signed up to use Meta’s apps, they signed terms and conditions that set out the balance of rights and responsibilities between them and the platforms they use. In short, they agreed to abide by the Meta’s rules, and in return Meta agreed to enforce those rules in order to make its platforms safe and useful.

Meta’s January 7th announcement makes radical changes to those rules, without even adequately explaining to users what changes to expect. That robs users of one of the few powers they can exercise, which is leaving a platform if they disagree with its rules.

As this report explains, there are huge questions about the scope and impact of Meta’s changes, with some described in just a few sentences in Meta’s blogpost, leaving huge ambiguities. Meta owes it to its users to answer these questions.

1 Meta will halt “proactive” enforcement of some policies on harmful content

- Meta will halt “proactive” enforcement of some policies on harmful content
- Instead, Meta will only act on relevant content after users report it
- Meta could drop millions of enforcement actions on harmful content
- Meta must clarify the scope of this change and how it will mitigate its impact

What Meta’s announcement says

Despite the significance of this change, Meta’s announcement describes it in just three sentences, leaving a lot of ambiguity about how it will be applied:

“Up until now, we have been using automated systems to scan for all policy violations, but this has resulted in too many mistakes and too much content being censored that shouldn’t have been. So, we’re going to continue to focus these systems on tackling illegal and high-severity violations, like terrorism, child sexual exploitation, drugs, fraud and scams. For less severe policy violations, we’re going to rely on someone reporting an issue before we take any action.”⁹

What Meta’s announcement means

Meta regularly releases transparency reports on the actions it takes against content violating its policies.¹⁰ It states that these enforcement actions include removing content, disabling accounts or “covering” harmful content such as violent images.¹¹

These reports distinguish between enforcement actions taken in response to user reports, and all other “proactive” enforcement in which Meta identifies and actions content itself, often using machine learning techniques to identify harmful content.¹²

Meta has previously credited this “proactive detection technology” as a key factor in reducing the prevalence of hate speech and other harmful content on its platforms.¹³ Mark Zuckerberg himself wrote in 2018 that this approach was “[the] single most important improvement in enforcing our policies”.¹⁴

Meta now says “proactive” enforcement will be halted for what it calls “less severe policy violations”, which it contrasts with “illegal and high-severity violations”.¹⁵

Given that the vast majority of Meta’s enforcement actions are “proactive”, the effect of this will be to radically cut Meta’s enforcement for “less severe policy violations”. Meta does not list some key policy areas, such as hate speech, amongst the short list of “illegal and high-severity violations” where proactive enforcement will continue.

Meta is not clear which policies will no longer receive proactive enforcement

Meta's announcement lists "illegal and high-severity violations" that will continue to be subject to proactive enforcement, including "terrorism, child sexual exploitation, drugs, fraud and scams."¹⁶

Meta is less clear about the policy areas where proactive enforcement will be halted. The company states that it will "rely on someone reporting an issue before we take any action" for "less severe policy violations", but it does not give examples of which policy areas may be affected by this change.

Assessing the scope of Meta's announcement

We can clarify the potential scope of Meta's announcement by comparing its recent announcement with the policy areas it defines in its regular Community Standards Enforcement Reports.¹⁷

Meta's announcement states that proactive enforcement will continue for "terrorism, child sexual exploitation, drugs, fraud and scams."¹⁸ In the table below, we have matched these categories with ones used in Meta's Community Standards Enforcement Reports. It is unclear if Meta intends for this list to be exhaustive.

This leaves a number of "policy areas" covered by Meta's Enforcement Reports that are not mentioned in Meta's latest announcement, making it unclear whether proactive enforcement will continue.¹⁹ These include important categories of harmful content, such as hate speech, incitement and bullying.

Meta must clarify the policy areas in which it will suspend proactive enforcement, in whole or in part. The following section attempts to estimate the impact of Meta halting proactive enforcement for "at risk" policy areas on the right side of the table below.

Proactive Enforcement Continues	Proactive Enforcement Unclear
Child Endangerment: Nudity and Physical Abuse	Adult Nudity & Sexual Activity ²⁰
Child Endangerment: Sexual Exploitation	Bullying & Harassment
Dangerous Orgs: Terrorism	Dangerous Orgs: Organized Hate
Fake Accounts	Hate Speech
Regulated Goods: Drugs	Suicide and Self-Injury
Regulated Goods: Firearms ²¹	Violence And Incitement
Spam	Violence & Graphic Content

Meta could halt 97% of its enforcement in key policy areas such as hate speech

We can assess the potential impact of Meta's decision to halt proactive enforcement for some policy areas by examining the company's transparency reports.²²

Our analysis shows that across policy areas where Meta could halt proactive enforcement (see previous section), over 97% of Meta's enforcement actions were "proactive" last year, with less than 3% made in response to user reports.

How we estimated Meta's proactive enforcement in key policy areas

Meta's Community Standards Enforcement Reports detail how many pieces of content Meta took enforcement action on globally, with figures available for each policy area and for each of its social media platforms, Facebook and Instagram. The company also details the proportion of these decisions that were a result of its proactive enforcement, in contrast to actions that were taken in response to user reports.²³

This allows us to calculate the number of proactive enforcement actions taken in each policy area where Meta may halt proactive enforcement (see previous section).²⁴ This analysis covers the last full year for which Meta has provided data as part of its quarterly Community Standards Enforcement Reports, running from the start of Q4 2023 to the end of Q3 2024.²⁵

Even excluding "mistakes", Meta correctly took action 277 million times last year

Meta has stated that "one to two out of every 10 [enforcement actions] may have been mistakes".²⁶ To arrive at a conservative estimate of **correct** enforcement actions, we have assumed that just 80% of all proactive enforcement actions are correct.

Even on this extremely conservative view of the accuracy of Meta's proactive enforcement, we estimate that Meta took proactive action against 277 million pieces of content **correctly** identified as harmful last year. This covers policy areas where Meta may halt proactive enforcement such as hate speech (see previous section).

Meta only provides global enforcement figures

Meta only provides global figures in its Community Standards Enforcement Reports.²⁷ These figures should therefore be taken as indicative of the potential impact of Meta's announcement if its changes to proactive enforcement were made worldwide.

Meta's announcement states that some of its changes, such as those to fact-checking, will be made "starting in the US".²⁸ At present, it is unclear how widely Meta plans to roll out its changes to proactive enforcement.

Table: Meta’s proactive enforcement from Q4 2023 to Q3 2024

Policy Area	Actions ²⁹	% Proactive ³⁰	# Proactive ³¹	#Proactive Accurate ³²
Bullying & Harassment	68,700,000	91.76%	63,040,200	50,432,160
Dangerous Orgs: Organized Hate	4,567,500	93.00%	4,248,000	3,398,400
Hate Speech	60,100,000	96.72%	58,128,600	46,502,880
Suicide and Self-Injury	49,700,000	99.34%	49,370,300	39,496,240
Violence And Incitement	71,500,000	98.76%	70,616,100	56,492,880
Violent & Graphic Content	101,700,000	98.97%	100,654,500	80,523,600
Total	356,267,500	97.13%	346,057,700	276,846,160

Questions to answer on proactive enforcement

Meta’s decision to communicate a wholesale change in its approach to enforcement with such scant detail makes it difficult to fully assess its impact. Meta should be pressed to answer the following questions in order to facilitate an informed public discussion about a decision that could affect hundreds of millions of users.

- Which of the policy areas listed in Meta’s Community Standards Enforcement Reports will be considered “less severe policy violations”, and therefore no longer subject to proactive enforcement?
- Which other policy areas, not listed in Meta’s Community Standards Enforcement Reports but previously subject to “proactive enforcement”, will no longer see proactive enforcement carried out?
- Meta claims to remove “misinformation [that] has the potential to cause imminent physical harm” or “has the potential to interfere with or suppress voting”, but does not provide transparency information on content actioned under this policy.³³ Does Meta propose to halt proactive enforcement for harmful misinformation too, and if so, how many actions against harmful disinformation will no longer take place?
- When does Meta plan to implement these changes to enforcement in the US?
- Meta has said it plans to expand these policy changes to countries beyond the US. Can it confirm which countries this is planned for, and when?

- Of the proactive enforcement actions that Meta reported taking in each area in the last year of its Community Standards Enforcement Reports, what proportion targeted content or accounts created by users in the US, UK or the EU?
- Analysis suggests that under its current policies, Meta took proactive enforcement action against 277 million pieces of content accurately assessed as harmful in the last year for which global data is available. How does it propose to manage the harm this content could pose to its users if it no longer takes these actions in future?
- Meta has previously credited its “proactive detection technology” as a key factor in reducing the prevalence of hate speech and other harmful content on the platform.³⁴ How does it propose to manage harm to its users if it ceases proactive enforcement and prevalence of harmful content rises as a result?
- Previous studies have shown that Meta appears to take enforcement action in response to a small minority of user reports of harmful content, including hate and potentially harmful disinformation.³⁵ If Meta proposes to rely primarily on user reports to enforce some policy areas, how will it ensure that those user reports are being dealt with quickly and accurately to manage harm?
- Meta claims that up to 20% of its proactive enforcement actions could have been mistakes. Can it provide data on how many mistakes it made in its proactive enforcement for different policy areas?

2 Meta will demote less content “that might violate our standards”

- Meta will demote less “content... that might violate our standards”
- Meta appears to be referring to what it previously called “borderline content”
- Mark Zuckerberg argued for demotion of “borderline content” in a 2018 blog
- He called this an “effective strategy” to reduce viral misinformation

What Meta’s announcement says

Meta states that it will be “getting rid” of most of the systems it presently uses to predict content that might violate its standards and demote it in users’ feeds:

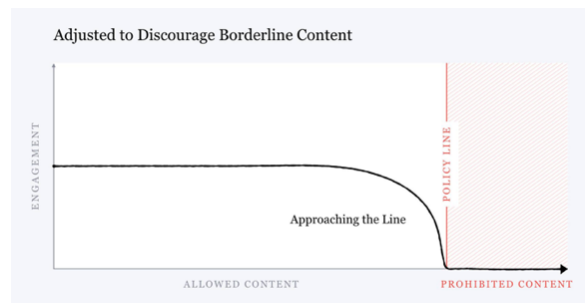
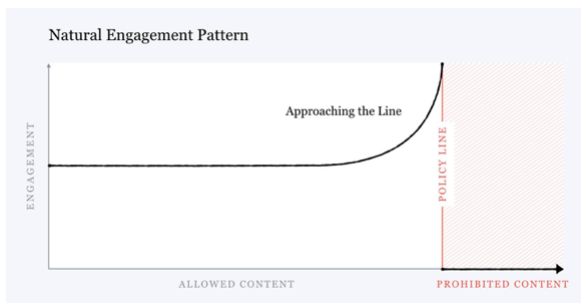
“We also demote too much content that our systems predict might violate our standards. We are in the process of getting rid of most of these demotions and requiring greater confidence that the content violates for the rest.”³⁶

What Meta’s announcement means

Meta has not set out the full set of demotions it will be removing, but the Platformer newsletter has reported that it will include demotions for misinformation.³⁷ Meta originally introduced demotions for misinformation in 2018.³⁸

In 2018, Mark Zuckerberg wrote that Meta’s own research found that the closer “borderline content” came to violating Meta’s policies, the more engagement it generated, creating a challenge for Meta’s enforcement teams.³⁹

He announced that Meta would use “AI systems to detect borderline content so we can distribute that content less”, in other words for targeted demotions.⁴⁰ He argued not only that this was an “effective strategy” for reducing the spread of misinformation, but that “[b]y fixing this incentive problem in our services, we believe it’ll create a virtuous cycle: by reducing sensationalism of all forms, we’ll create a healthier, less polarized discourse where more people feel safe participating.”⁴¹



Charts used by Mark Zuckerberg to illustrate how content attracts more engagement the closer it is to violating platform policies (left), and the counteracting effect of demoting borderline content (right)

Questions to answer on content demotions

- Mark Zuckerberg previously argued that demoting “borderline content” was the “next most effective strategy” to reduce the spread of misinformation, after removing fake accounts that generate it. Why is Meta dropping a strategy it says helps reduce misinformation, a type of content that “people consistently tell us... makes our service worse”?⁴²
- Mark Zuckerberg previously argued that demoting “borderline content” would help “create a healthier, less polarized discourse where more people feel safe participating.”⁴³ What assessment has Meta made of the impact of removing these demotions?

3 Meta is dropping policies on “immigration, gender identity and gender”

- Meta will drop policies on “immigration, gender identity and gender”
- Meta will allow more “insulting language” about immigration and gender
- Guidance issued to Meta moderators shows what content is now permitted
- Meta’s policy changes “have been implemented worldwide immediately”

What Meta’s announcement says

Meta announced that it will reduce policy restrictions on “immigration, gender identity and gender”, and has published an updated hateful conduct policy reflecting this:⁴⁴

“We’re getting rid of a number of restrictions on topics like immigration, gender identity and gender that are the subject of frequent political discourse and debate. [...] These policy changes may take a few weeks to be fully implemented.”⁴⁵

What Meta’s announcement means

Journalists have obtained new training materials issued to Meta’s content moderators that make the full extent of these policy changes clear.⁴⁶ These documents explain that the following statements on gender are now permitted under Meta’s new policies:

- “Trans people aren’t real. They’re mentally ill.”
- “Gays are not normal.”
- “Women are crazy.”
- “A trans person isn’t a he or she, it’s an it.”⁴⁷
- “Trannies are a problem.”⁴⁸

The same documents show that Meta’s announcement leaves much more room for claims about migrants and racial groups.⁴⁹ Acceptable statements reportedly include:

- “Immigrants are grubby, filthy pieces of shit.”
- “These damn immigrants can’t be trusted, they’re all criminals.”
- “Black people are more violent than Whites.”
- “Mexicans are lazier than Asians.”
- “Jews are flat out greedier than Christians.”⁵⁰

The Intercept notes that the guidance issued to moderators appears contradictory or convoluted in places, for example allowing “trans people are mentally ill” while banning “lesbians are so stupid”, or allowing “migrants are no better than vomit” while banning “Muslims make me want to throw up”.⁵¹ This may hinder effective enforcement.

Meta’s Chief Global Affairs Officer, Joel Kaplan, has said that these changes to Meta’s hateful conduct policies “have been implemented worldwide immediately.”⁵²

Questions to answer on policy changes

Meta should be pressed to answer the following questions in order to facilitate an informed public discussion about policy changes that could affect millions of users.

- What assessment has Meta made of the impact that these policy changes could have on the safety of its users?
- Has Meta consulted with groups who could be affected by these changes so that the impact on them is understood?
- What is Meta's rationale for issuing internal guidance instructing moderators to allow "statements of superiority as long as the statements do not refer to inferiority of another [protected characteristic] group (a) on the basis of inherent intellectual ability and (b) without support"?⁵³
- What is Meta's rationale for instructing moderators that it is permissible to post statements like "Black people are more violent than Whites", "Mexicans are lazier than Asians" or "Jews are flat out greedier than Christians"?⁵⁴

4 Meta will replace independent fact-checking with Community Notes

- Meta will stop attaching independent fact-checks to false or misleading posts
- It will instead use Community Notes, a system of user-generated fact-checks
- Existing Community Notes tools perform poorly on important, divisive topics
- There are obstacles to introducing Community Notes effectively on Facebook

Meta plans to replace independent fact-checking labels on posts in the coming months, replacing it with a system of user-generated fact-checks. This new policy will initially roll out in the US, but Meta says it will expand to other countries later.⁵⁵

What Meta's announcement says

Meta has announced its intention to replace independent fact-checking labels with a system of user-generated fact-checks, modelled on X's Community Notes feature:

We will end the current third party fact checking program in the United States and instead begin moving to a Community Notes program. [...] Just like they do on X, Community Notes will require agreement between people with a range of perspectives to help prevent biased ratings. We intend to be transparent about how different viewpoints inform the Notes displayed in our apps, and are working on the right way to share this information. [...] We plan to phase in Community Notes in the US first over the next couple of months, and will continue to improve it over the course of the year.⁵⁶

What Meta's announcement means

Meta has contributed \$100 million to independent fact-checkers since partnering with them in 2016.⁵⁷ Content rated as false or misleading by fact-checkers is labelled, demoted in news feeds and displays a warning to users who share it.⁵⁸

Meta says it will terminate this program in the US in the coming months.⁵⁹ In its place, Meta plans to replicate Community Notes, X's system for providing user-generated fact-checks or additional context on posts.⁶⁰ In short, the system works by having users write and rate proposed labels for posts, only displaying them to all users if they receive positive ratings from users with "diverse perspectives".⁶¹

Meta has indicated that it will expand its rollout of Community Notes to other countries once it has been implemented and improved in the US, with the company's Chief Global Affairs Officer Joel Kaplan stated they expect this to start happening in 2026.⁶²

Meta has previously said that it removes "misinformation [that] has the potential to cause imminent physical harm" or "has the potential to interfere with or suppress voting."⁶³ It is not clear if Meta's latest announcement changes this.

Community Notes perform poorly on polarizing, controversial topics

While Community Notes have won praise for enabling users to address the problem of false and misleading content, they have significant weaknesses.⁶⁴

Studies indicate that Community Notes perform poorly when it comes to divisive topics where it's rare for contributors from a range of political leanings to reach consensus on whether a note is "Helpful".⁶⁵ Our research has repeatedly found similar weaknesses:

- 74% of accurate Community Notes that fact-checked misinformation in the US elections did not reach the rating consensus needed to be shown to all users.⁶⁶
- None of Elon Musk's posts promoting misleading claims about the US elections received a Community Note, despite the posts amassing 2 billion views.⁶⁷
- Just one out of 1,060 posts promoting misleading claims that contributed to riots targeting migrants and Muslims in the UK displayed a Community Note.⁶⁸

In addition, we found that even in cases where misleading posts about elections displayed Community Notes on X, those posts had been viewed 13 times more than the notes on them.⁶⁹ Other research has found a similar trend across topics, with Community Notes viewed far less than the posts they are applied to.⁷⁰

For these reasons, we should not expect Meta's implementation of Community Notes to serve as a true replacement for independent fact-checks, particularly for disinformation on polarizing issues or for important events like elections.

Meta faces challenges in implementing Community Notes on Facebook

For a Community Notes system to be effective, a wide range of users must be able to find posts they might disagree with and rate draft Community Notes on them.

This is relatively easy on X, where posts are public and visible to all users by default. But it may prove more challenging on Facebook, where many posts exist inside Groups and are only visible to their members, who are likely to share similar views. Meta states that "[m]ore than 1.8 billion people use Facebook Groups every month, and more than half are in five or more groups".⁷¹

For example, our previous research identified Facebook Groups that not only host antisemitic content but are operated by antisemitic admins.⁷² Content posted in such groups is unlikely to be seen by users with dissenting views – indeed, we found that these antisemitic admins would sometimes remove users who report hateful content.

Groups are an integral part of Facebook, with some groups containing millions of members. Meta must therefore explain if and how its Community Notes system will work effectively for content posted into Groups.

Questions to answer on fact-checking

Meta has announced an end to its fact-checking program, while saying little about how it will implement a Community Notes system that addresses similar problems. Meta should be pressed to answer the following questions to facilitate an informed public discussion about a decision that could affect hundreds of millions of users.

- Meta has stated an intention to replace independent fact-checks with a Community System in countries beyond the US. Which non-US markets does Meta plan to do this in, and when?
- Meta would previously demote content that promoted clearly false or misleading hoaxes. Does it plan to do the same for posts identified as false or misleading by Community Notes?
- Existing research suggests that Community Notes perform poorly on important but controversial topics where it is hard for a wide range of users to reach consensus. How has Meta assessed the risk this could pose in terms of the spread of harmful misleading or false content that no longer carries a fact-checking label?
- How does Meta propose to implement Community Notes in a way that enables a wide range of users to propose and rate additional context for popular content posted inside Facebook Groups?
- Meta's existing policies state that it will remove "misinformation [that] has the potential to cause imminent physical harm" or "has the potential to interfere with or suppress voting."⁷³ How will Meta determine which posts violate this policy in the absence of independent fact-checks?

5 Meta will demote less content about “elections, politics or social issues”

- Meta will show users more content about “elections, politics or social issues”
- This is a reversal of Meta’s 2021 decision to show users less of this content
- Meta previously said users wanted to see less political content in their feeds
- Leaked documents reveal concerns that Meta was promoting “civic disinfo”

What Meta’s announcement says

Meta will reverse its 2021 decision to reduce the amount of content about “elections, politics or social issues” users see, instead treating this like any other content.⁷⁴

“...we’re going to start treating civic content from people and Pages you follow on Facebook more like any other content in your feed, and we will start ranking and showing you that content based on explicit signals (for example, liking a piece of content) and implicit signals (like viewing posts) that help us predict what’s meaningful to people. We are also going to recommend more political content based on these personalized signals and are expanding the options people have to control how much of this content they see.”⁷⁵

What Meta’s announcement means

In 2021, Mark Zuckerberg announced Meta would reduce the amount of political content that users see in their News Feeds, building on a decision to “[stop] recommending civil and political groups” in the 2020 US elections.⁷⁶

At the time, Mark Zuckerberg said these changes were a response to user feedback that “people don’t want politics and fighting to take over their experience on our services.”⁷⁷

Indeed, documents leaked by the whistleblower Frances Haugen reveal that around this time Meta’s own researchers were finding that “civic content is more likely to be rated [bad for the world] than non-civic content” in surveys of Facebook users.⁷⁸

Other documents suggest that Facebook’s prioritization of posts that generate lots of user interactions was thought to be “contributing hugely to Civic misinfo”.⁷⁹

Despite this history, Meta has effectively reversed its position on civic content and intends to rank and recommend it much like any other content on its platforms.⁸⁰

Questions to answer on civic content

- Why is Meta taking this moment to show users more civic content that it previously claimed users consider ‘bad for the world’ and do not want to see?

6 Meta claims it will move trust and safety teams from California to Texas

- Meta will move trust & safety teams from California to states including Texas
- Mark Zuckerberg says this means “less concern about the bias of our teams”
- Meta also says that it will prioritize user appeals against enforcement actions
- It is unclear what actual impact this will have on Meta’s trust & safety staffing

What Meta’s announcement says

Meta says it will move trust and safety teams from California to Texas and other states, as well as investing more staff time in appeals against enforcement actions:

“As part of these changes, we will be moving the trust and safety teams that write our content policies and review content out of California to Texas and other US locations. People are often given the chance to appeal our enforcement decisions and ask us to take another look, but the process can be frustratingly slow and doesn’t always get to the right outcome. We’ve added extra staff to this work and in more cases, we are also now requiring multiple reviewers to reach a determination in order to take something down.”⁸¹

What Meta’s announcement means

In the video released to accompany Meta’s policy announcements, Mark Zuckerberg claims that “our US-based content review is going to be based in Texas”, arguing “as we work to promote free expression, I think it will help us build trust to do this work in places where there’s less concern about the bias of our teams.”⁸²

However, it is important to note that Meta reportedly already has “major content moderation and trust and safety operations” in Texas.⁸³ Meta has so far declined to answer questions from reporters on the nature and scope of these changes.⁸⁴ Other reports have suggested that the announcement would affect only a small number of staff working for Meta in California.⁸⁵

Meta does not report on how many content moderation staff it has in each state, making the overall impact of these changes unclear.

Questions to answer on content moderation teams

- Will Meta maintain, increase or decrease staffing on its US trust and safety teams in the coming year?
- Given that Meta is planning to rely more heavily on user reports for some policy violations, will Meta increase content moderation teams to handle this?

References

¹ “An Independent Assessment of the Human Rights Impact of Facebook in Myanmar”, Meta, 5 November 2018, <https://about.fb.com/news/2018/11/myanmar-hria/>

² “Inside Meta’s dehumanizing new speech policies for trans people”, Platformer, 9 January 2025, <https://www.platformer.news/meta-new-trans-guidelines-hate-speech/>

“Leaked Meta Rules: Users are free to post ‘Mexican immigrants are trash!’ or ‘Trans people are immoral’”, The Intercept, 9 January 2025, <https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/>

³ “More Speech and Fewer Mistakes”, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

⁴ “More Speech and Fewer Mistakes”, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

Mark Zuckerberg, Threads, 7 January 2025, <https://www.threads.net/@zuck/post/DEhgYx4JbEG>

⁵ 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/>

⁶ 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/>

⁷ “Meta Earnings Presentation Q4 2024”, Meta, 29 January 2025,

<https://investor.atmeta.com/financials/default.aspx>

⁸ “Social Media Fact Sheet”, Pew Research Center, 13 November 2024,

<https://www.pewresearch.org/internet/fact-sheet/social-media/>

⁹ “More Speech and Fewer Mistakes”, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

¹⁰ “Community Standards Enforcement Report”, Meta, retrieved 28 January 2025,

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

¹¹ "Taking action could include removing a piece of content from Facebook or Instagram, covering photos or videos that may be disturbing to some audiences with a warning, or disabling accounts. In the event that we escalate content to law enforcement, we don't additionally count that."

"Content actioned", Meta, 7 November 2023, <https://transparency.meta.com/en-gb/policies/improving/content-actioned-metric/>

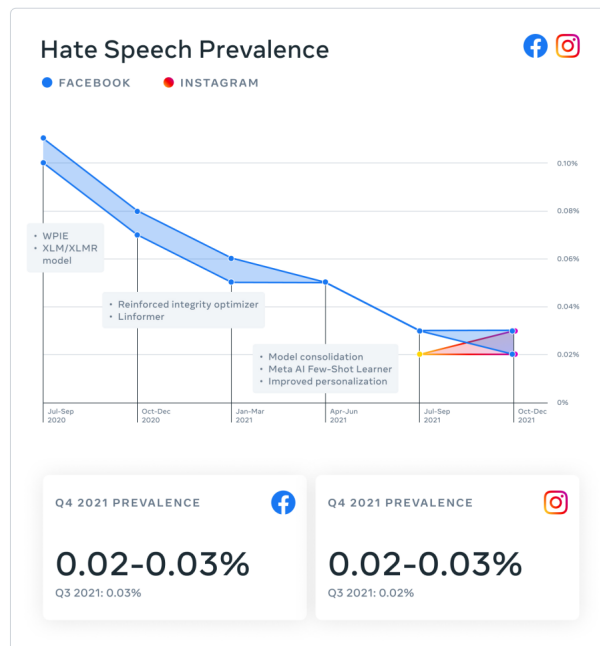
¹² "Community Standards Enforcement Report", Meta, retrieved 28 January 2025, <https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/>

"Proactive rate", Meta, 22 February 2023, <https://transparency.meta.com/en-gb/policies/improving/proactive-rate-metric/>

¹³ "Prevalence of hate speech content was about 0.03% in Q3 2021, which marks a decrease from Q2 2021 due to our proactive detection technology and ranking changes to personalize content for users and reduce problematic content in News Feed."

"Safety and integrity quarterly roundup", Meta, 10 November 2021, [link](#)

"With advancements in AI technologies, like our Few Shot Learner and shift to generalized AI, we've been able to correctly take action on harmful content more quickly."



"Community Standards Enforcement Report, Fourth Quarter 2021", Meta, 1 March 2022, <https://about.fb.com/news/2022/03/community-standards-enforcement-report-q4-2021/>

¹⁴ "A Blueprint for Content Governance and Enforcement", Mark Zuckerberg, Facebook, 15 November 2018, <https://www.facebook.com/notes/751449002072082/>

¹⁵ "More Speech and Fewer Mistakes", Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

¹⁶ It is unclear if the policy areas that Meta lists in the announcement are intended to be exhaustive or merely illustrative examples of the kind of "illegal and high-severity violations" where it will continue to carry out proactive enforcement.

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

¹⁷ See "Data by policy area" in Meta's latest Community Standards Enforcement Report.

“Community Standards Enforcement Report”, Meta, retrieved 4 February 2025, <https://transparency.meta.com/reports/community-standards-enforcement/>

¹⁸ “More Speech and Fewer Mistakes”, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

¹⁹ See “Data by policy area” in Meta’s latest Community Standards Enforcement Report.

“Community Standards Enforcement Report”, Meta, retrieved 4 February 2025, <https://transparency.meta.com/reports/community-standards-enforcement/>

²⁰ We have excluded Adult Nudity & Sexual Activity from analysis in the following section as it is unclear if content in this policy area can be described as “harmful” in the same way as other policy areas.

²¹ While Meta’s announcement does not explicitly refer to firearms in the same way it does drugs, its transparency reports include both in the broader category of “regulated goods”. We have therefore interpreted firearms as an area in which Meta will continue to enforce “illegal” violations.

²² “Community Standards Enforcement Report”, Meta, retrieved 29 January 2025, <https://transparency.meta.com/reports/community-standards-enforcement/>

²³ “Proactive rate”, Meta, 22 February 2023, <https://transparency.meta.com/en-gb/policies/improving/proactive-rate-metric/>

²⁴ “Community Standards Enforcement Report”, Meta, retrieved 29 January 2025, <https://transparency.meta.com/reports/community-standards-enforcement/>

²⁵ “Community Standards Enforcement Report”, Meta, retrieved 29 January 2025, <https://transparency.meta.com/reports/community-standards-enforcement/>

²⁶ “For example, in December 2024, we removed millions of pieces of content every day. While these actions account for less than 1% of content produced every day, we think one to two out of every 10 of these actions may have been mistakes (i.e., the content may not have actually violated our policies). This does not account for actions we take to tackle large-scale adversarial spam attacks.”

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

²⁷ “Community Standards Enforcement Report”, Meta, retrieved 29 January 2025, <https://transparency.meta.com/reports/community-standards-enforcement/>

²⁸ “Starting in the US, we are ending our third party

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

²⁹ This is the sum of total enforcement actions that Meta took under each policy area as reported in its Q4 2023, Q1 2024, Q2 2024 and Q3 2024 Community Standards Enforcement Reports.

“Community Standards Enforcement Report”, Meta, retrieved 29 January 2025, <https://transparency.meta.com/reports/community-standards-enforcement/>

³⁰ Meta defines the “proactive rate” metric in its quarterly Community Standards Enforcement Reports as “the percentage of all content or accounts acted on that we found and actioned before users reported them to us.” We have used Meta’s quarterly data on enforcement actions and proactive rate in each of Q4 2023, Q1 2024, Q2 2024 and Q3 2024 to calculate the effective proactive rate across the whole year under analysis.

“Proactive rate”, Meta, 22 February 2023, <https://transparency.meta.com/en-us/policies/improving/proactive-rate-metric/>

³¹ This is the total number of proactive enforcement actions under each policy area reported by Meta in its quarterly Community Standards Enforcement Reports, distinguished from actions taken in response to user reports. For each quarter, we multiplied the number of enforcement actions by the percentage of those actions that were proactive. We then summed these quarterly values to reach a total number of proactive enforcement actions for the year spanning Q4 2023, Q1 2024, Q2 2024 and Q3 2024.

"Community Standards Enforcement Report", Meta, retrieved 29 January 2025, <https://transparency.meta.com/reports/community-standards-enforcement/>

³² Meta has stated in its latest announcement that "one to two out of every 10 [enforcement actions] may have been mistakes". To arrive at a conservative estimate of accurate enforcement actions, we have assumed that just 80% of all proactive enforcement actions are correct.

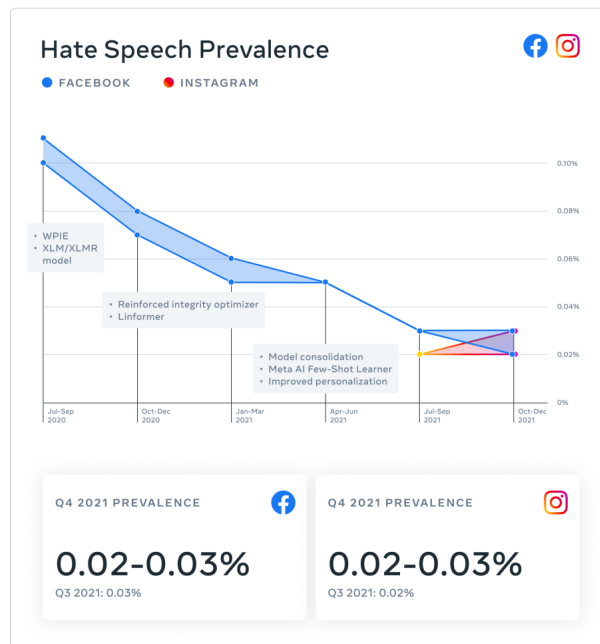
"More Speech and Fewer Mistakes", Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

³³ "Our approach to misinformation", 22 January 2025, <https://transparency.meta.com/en-gb/features/approach-to-misinformation/>

³⁴ "Prevalence of hate speech content was about 0.03% in Q3 2021, which marks a decrease from Q2 2021 due to our proactive detection technology and ranking changes to personalize content for users and reduce problematic content in News Feed."

"Safety and integrity quarterly roundup", Meta, 10 November 2021, [link](#)

"With advancements in AI technologies, like our Few Shot Learner and shift to generalized AI, we've been able to correctly take action on harmful content more quickly."



"Community Standards Enforcement Report, Fourth Quarter 2021", Meta, 1 March 2022, <https://about.fb.com/news/2022/03/community-standards-enforcement-report-q4-2021/>

³⁵ "Failure to Act", Center for Countering Digital Hate, 1 December 2020, <https://counterhate.com/research/failure-to-act/>

"Failure to Protect", Center for Countering Digital Hate, 30 July 2021, <https://counterhate.com/research/failure-to-protect/>

"Hidden Hate", Center for Countering Digital Hate, 6 April 2022, <https://counterhate.com/research/hidden-hate/>

"Failure to Protect", Center for Countering Digital Hate, 28 April 2022, <https://counterhate.com/research/anti-muslim-hate/>

"Antisemitic Admins", Center for Countering Digital Hate, 29 May 2024, <https://counterhate.com/research/antisemitic-admins/>

"Abusing Women in Politics", Center for Countering Digital Hate, 14 August 2024, <https://counterhate.com/research/abusing-women-in-politics/>

³⁶ "More Speech and Fewer Mistakes", Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

³⁷ "Meta just flipped the switch that prevents misinformation from spreading in the United States", Platformer, 14 January 2025, <https://www.platformer.news/meta-ends-misinformation-enforcement-zuckerberg/>

³⁸ "How People Help Fight False News", Meta, 21 June 2018, <https://about.fb.com/news/2018/06/inside-feed-how-people-help-fight-false-news/>

³⁹ "A Blueprint for Content Governance and Enforcement", Mark Zuckerberg, Facebook, 15 November 2018, <https://www.facebook.com/notes/751449002072082/>

⁴⁰ "This process for adjusting this curve is similar to what I described above for proactively identifying harmful content, but is now focused on identifying borderline content instead. We train AI systems to detect borderline content so we can distribute that content less."

"A Blueprint for Content Governance and Enforcement", Mark Zuckerberg, Facebook, 15 November 2018, <https://www.facebook.com/notes/751449002072082/>

⁴¹ "The category we're most focused on is click-bait and misinformation. People consistently tell us these types of content make our services worse -- even though they engage with them. As I mentioned above, the most effective way to stop the spread of misinformation is to remove the fake accounts that generate it. The next most effective strategy is reducing its distribution and virality. [...] I believe these efforts on the underlying incentives in our systems are some of the most important work we're doing across the company. We've made significant progress in the last year, but we still have a lot of work ahead. By fixing this incentive problem in our services, we believe it'll create a virtuous cycle: by reducing sensationalism of all forms, we'll create a healthier, less polarized discourse where more people feel safe participating."

"A Blueprint for Content Governance and Enforcement", Mark Zuckerberg, Facebook, 15 November 2018, <https://www.facebook.com/notes/751449002072082/>

⁴² "The category we're most focused on is click-bait and misinformation. People consistently tell us these types of content make our services worse -- even though they engage with them. As I mentioned above, the most effective way to stop the spread of misinformation is to remove the fake accounts that generate it. The next most effective strategy is reducing its distribution and virality."

"A Blueprint for Content Governance and Enforcement", Mark Zuckerberg, Facebook, 15 November 2018, <https://www.facebook.com/notes/751449002072082/>

⁴³ "I believe these efforts on the underlying incentives in our systems are some of the most important work we're doing across the company. We've made significant progress in the last year, but we still have a lot of work ahead. By fixing this incentive problem in our services, we believe it'll create a virtuous cycle: by reducing sensationalism of all forms, we'll create a healthier, less polarized discourse where more people feel safe participating."

"A Blueprint for Content Governance and Enforcement", Mark Zuckerberg, Facebook, 15 November 2018, <https://www.facebook.com/notes/751449002072082/>

⁴⁴ "Hateful Conduct", Meta, 7 January 2025, <https://transparency.meta.com/en-us/policies/community-standards/hate-speech/>

⁴⁵ "More Speech and Fewer Mistakes", Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

⁴⁶ "Inside Meta's dehumanizing new speech policies for trans people", Platformer, 9 January 2025, <https://www.platformer.news/meta-new-trans-guidelines-hate-speech/>

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

⁴⁷ "Inside Meta's dehumanizing new speech policies for trans people", Platformer, 9 January 2025, <https://www.platformer.news/meta-new-trans-guidelines-hate-speech/>

⁴⁸ "Leaked Meta Rules: Users are free to post 'Mexican immigrants are trash!' or 'Trans people are immoral'", The Intercept, 9 January 2025, <https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/>

⁴⁹ "Leaked Meta Rules: Users are free to post 'Mexican immigrants are trash!' or 'Trans people are immoral'", The Intercept, 9 January 2025, <https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/>

⁵⁰ "Leaked Meta Rules: Users are free to post 'Mexican immigrants are trash!' or 'Trans people are immoral'", The Intercept, 9 January 2025, <https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/>

⁵¹ "Leaked Meta Rules: Users are free to post 'Mexican immigrants are trash!' or 'Trans people are immoral'", The Intercept, 9 January 2025, <https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/>

⁵² 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/>

⁵³ "Leaked Meta Rules: Users are free to post 'Mexican immigrants are trash!' or 'Trans people are immoral'", The Intercept, 9 January 2025, <https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/>

⁵⁴ "Leaked Meta Rules: Users are free to post 'Mexican immigrants are trash!' or 'Trans people are immoral'", The Intercept, 9 January 2025, <https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/>

⁵⁵ "Community Notes are beginning to be rolled out in the US and will continue to improve it over the course of the year before expansion to other countries (read more here)."

"Our approach to misinformation", Meta, 22 January 2025, <https://transparency.meta.com/en-gb/features/approach-to-misinformation/>

⁵⁶ "More Speech and Fewer Mistakes", Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

⁵⁷ "Meta's Investments in Fact-Checking", Meta, 1 April 2022, <https://www.facebook.com/formedia/blog/third-party-fact-checking-industry-investments>
Mark Zuckerberg, Facebook, 18 November 2016, <https://www.facebook.com/zuck/posts/10103269806149061>

"Addressing Hoaxes and Fake News", Meta, 15 December 2016, <https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/>

⁵⁸ “Our approach to misinformation”, Meta, 22 January 2025, <https://transparency.meta.com/en-gb/features/approach-to-misinformation/>

⁵⁹ “How fact-checking works”, Meta, 7 January 2025, <https://transparency.meta.com/en-gb/features/how-fact-checking-works/>

⁶⁰ “More Speech and Fewer Mistakes”, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

⁶¹ “Community Notes: a collaborative way to add helpful context to posts and keep people better informed”, Meta, retrieved 30 January 2025, <https://communitynotes.x.com/guide/en/about/introduction>

⁶² “Community Notes are beginning to be rolled out in the US and will continue to improve it over the course of the year before expansion to other countries (read more here).”

“Our approach to misinformation”, Meta, 22 January 2025, <https://transparency.meta.com/en-gb/features/approach-to-misinformation/>

Joel Kaplan: “So we’ve decided to move away from a third-party fact-checking program, starting in the US, and to replace it with a Community Notes-based system, really a crowd-sourced system, to take advantage of the broad, diverse set of views that we have represented by the people who use our platform. And we’re going to be rolling that out over the next couple of months. As I said starting the US, we’re going to build it there, iterate there, see what works, see what improvements we need to make before we launch elsewhere, probably starting in 2026.”

“Joel Kaplan on EU Regulation and Innovation”, Meta, 02:35, 7 February 2025, <https://about.fb.com/news/2025/02/joel-kaplan-on-eu-regulation-and-innovation/>

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/>

⁶³ “Our approach to misinformation”, Meta, 22 January 2025, <https://transparency.meta.com/en-gb/features/approach-to-misinformation/>

⁶⁴ “Building a better Birdwatch”, X, 3 March 2022, https://blog.x.com/en_us/topics/company/2022/building-a-better-birdwatch

⁶⁵ “Twitter’s Fact-Checking System Has a Major Blind Spot: Anything Divisive”, Bloomberg, 19 December 2022, <https://www.bloomberg.com/graphics/2022-twitter-birdwatch-community-notes-misinformation-politics/>

⁶⁶ “Rated Not Helpful”, Center for Countering Digital Hate, 30 October 2024, <https://counterhate.com/research/rated-not-helpful-x-community-notes/>

⁶⁷ “Musk’s Political Posts”, Center for Countering Digital Hate, 4 November 2024, <https://counterhate.com/research/musk-political-posts-x/>

⁶⁸ “X ran ads on five accounts pushing lies and hate during UK riots”, X, 19 August 2024, <https://counterhate.com/research/x-ran-ads-on-five-accounts-pushing-lies-and-hate-during-uk-riots/>

⁶⁹ “Rated Not Helpful”, Center for Countering Digital Hate, 30 October 2024,

<https://counterhate.com/research/rated-not-helpful-x-community-notes/>

⁷⁰ “Most users on X never see Community Notes correcting misinformation”, Mashable, 30 November 2023, <https://mashable.com/article/twitter-x-community-notes-misinformation-views-investigation>

⁷¹ “The Power of Virtual Communities”, Meta, 23 February 2021,

<https://www.facebook.com/community/whats-new/power-virtual-communities/>

⁷² “Antisemitic Admins”, Center for Countering Digital Hate, 29 May 2024,

<https://counterhate.com/research/antisemitic-admins/>

⁷³ “Our approach to misinformation”, Meta, 22 January 2025, <https://transparency.meta.com/en-gb/features/approach-to-misinformation/>

⁷⁴ “Since 2021, we’ve made changes to reduce the amount of civic content people see – posts about elections, politics or social issues – based on the feedback our users gave us that they wanted to see less of this content. But this was a pretty blunt approach. We are going to start phasing this back into Facebook, Instagram and Threads with a more personalized approach so that people who want to see more political content in their feeds can.”

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

“Reducing Political Content in News Feed”, Meta, 10 February 2021,

<https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/>

⁷⁵ “More Speech and Fewer Mistakes”, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

⁷⁶ “Along these same lines, we’re also currently considering steps we could take to reduce the amount of political content in News Feed as well. We’re still working through exactly the best ways to do this. And to be clear, of course we’ll still enable people to engage in political groups and discussions if they want to. These can often be important and helpful. They can be ways that people organize grassroots movements, speak out against injustice, or learn from people with different perspectives. So we want these discussions to be able to keep happening. But one of the top pieces of feedback we’re hearing 3 from our community right now is that people don’t want politics and fighting to take over their experience on our services.”

Mark Zuckerberg, “Fourth Quarter 2020 Results Conference Call”, Meta, 27 January 2021,

https://s21.q4cdn.com/399680738/files/doc_financials/2020/q4/FB-Q4-2020-Conference-Call-Transcript.pdf

⁷⁷ “Along these same lines, we’re also currently considering steps we could take to reduce the amount of political content in News Feed as well. We’re still working through exactly the best ways to do this. And to be clear, of course we’ll still enable people to engage in political groups and discussions if they want to. These can often be important and helpful. They can be ways that people organize grassroots movements, speak out against injustice, or learn from people with different perspectives. So we want these discussions to be able to keep happening. But one of the top pieces of feedback we’re hearing 3 from our community right now is that people don’t want politics and fighting to take over their experience on our services.”

Mark Zuckerberg, “Fourth Quarter 2020 Results Conference Call”, Meta, 27 January 2021,

https://s21.q4cdn.com/399680738/files/doc_financials/2020/q4/FB-Q4-2020-Conference-Call-Transcript.pdf

⁷⁸ “Facebook Asked Users What Content Was ‘Good’ or ‘Bad for the World.’ Some of the Results Were Shocking”, Slate, 15 November 2021, <https://slate.com/technology/2021/11/facebook-good-bad-for-the-world-gftw-bftw.html>

“How much of News Feed is Good (or Bad) for the world?”, Meta, 23 October 2020,

https://www.documentcloud.org/documents/21748444-tier2_fa_news_ir_1020/

⁷⁹ “In a product briefing called “Ranking For Civic Health,” Facebook acknowledged that “people think that political content on Facebook is low quality, untrustworthy, and divisive,” and the current ranking system was “not creating a wholly valuable civic experience for users.” (Based on document comment dates, the document was produced around January and February of 2020.)”

“Eight things we learned from the Facebook Papers”, The Verge, 25 October 2021,

<https://www.theverge.com/22740969/facebook-files-papers-frances-haugen-whistleblower-civic-integrity>

“Product Brief: Ranking for Civic Health”, Meta, https://www.documentcloud.org/documents/21091530-ranking-for-civic-health_sanitized/

⁸⁰ “As a result, **we’re going to start treating civic content from people and Pages you follow on Facebook more like any other content in your feed**, and we will start ranking and showing you that content based on explicit signals (for example, liking a piece of content) and implicit signals (like viewing posts) that help us predict what’s meaningful to people.”

“More Speech and Fewer Mistakes” (emphasis added), Meta, 7 January 2025,

<https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

⁸¹ “More Speech and Fewer Mistakes”, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

⁸² “More Speech and Fewer Mistakes”, Meta, 7 January 2025, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

⁸³ “Former Facebook employees say, however, that the move-to-Texas announcement rings hollow. That’s because Meta already has major content moderation and trust and safety operations in the state.”

“Meta moderators were already in Texas before Zuckerberg announced move, say ex-workers”, The Guardian, 13 January 2025, <https://www.theguardian.com/technology/2025/jan/13/meta-moderators-texas-zuckerberg-trump>

⁸⁴ “Meta has several offices in Texas, mostly clustered in Austin – a progressive city in the conservative state. The company, which owns Facebook, Instagram, WhatsApp and Threads, did not respond to multiple requests for comment. It did not answer questions about how many employees will be affected and if trust and safety teams are moving to Texas or simply “out of California”. Meta also didn’t respond to questions about whether the affected workers are full-time employees or contractors.”

“Meta moderators were already in Texas before Zuckerberg announced move, say ex-workers”, The Guardian, 13 January 2025, <https://www.theguardian.com/technology/2025/jan/13/meta-moderators-texas-zuckerberg-trump>

⁸⁵ “Not Many Meta Employees Will Have to Move to Texas After All”, Wired, 15 January 2025, <https://www.wired.com/story/not-many-meta-employees-move-to-texas-content-moderation/>



Published February 2025
© Center for Countering Digital Hate Inc