

# **GARM** Aggregated Measurement Report

**Volume 5** | August 2023



## Contents

<b>03</b>	Aggregated Measurement Report	<b>63</b>	TikTok
<b>04</b>	Using the GARM Aggregated Measurement Report	<b>75</b>	Pinterest
<b>06</b>	Executive Summary	<b>91</b>	Snapchat
<b>09</b>	YouTube	<b>100</b>	Twitch
<b>22</b>	Meta	<b>109</b>	LinkedIn
<b>28</b>	Facebook	<b>117</b>	Appendices & FAQ
<b>42</b>	Instagram		
<b>55</b>	Twitter		

# Creating the GARM Aggregated Measurement Report

In June 2019, we established the Global Alliance for Responsible Media (GARM) to create a more sustainable and responsible digital environment that protects consumers, the media industry, and society as a result.

**Since our launch, we've been focused on creating value for society and the advertising industry in three strategic focus areas:**

1. Establishing shared, common definitions on harmful content for advertising & media
2. Improving and creating common brand safety tools across the industry
3. Driving mutual accountability, and independent verification and oversight

The GARM Aggregated Measurement Report is our first solution in accountability. This report, like other GARM solutions, advances existing individual practices and establishes a common framework for better access, understanding, and for driving better practices.

## Why are we creating this report?

YouTube, Facebook, and Twitter all provided content policy reporting in 2018. Over time more digital media platforms have adopted this practice with the goal of communicating effective content moderation practices to several stakeholder audiences, ranging from regulators to NGOs to advertisers. With GARM's focus on societal safety and media industry sustainability, we want to more accurately communicate progress and challenges in individual and collective work to eliminate harmful content from ad supported media.

We've created the GARM Aggregated Measurement Report with advertising industry stakeholders in mind, and are delivering value through the following 5 steps:

### Creating a single access point

Our first step was to streamline access to data across platforms – we created a shared report with a year's worth of data from each platform that fundamentally improves access and visibility. In doing this, we've eliminated the need to extract data from individual period-based reports.

### Establishing a framework for industry focus

Our second step was to create a framework that creates focus on measures that should matter most to advertisers. We've done this based on a series of four core questions that we could rightly ask ourselves as an industry.

### Defining a set of quality metrics to answer critical questions

Our third step was to agree on measures that are best set up to answer the four core questions asked. This has resulted in the industry agreeing to best practices (authorized metrics), with an understanding that they would be pursued over time. In the absence of an authorized metric, a next best metric can be submitted by the platform so long as it helps to answer the question.

### Creating a link between policy to established categories

Our fourth step is to link existing platform policies reporting to the GARM Brand Safety Floor categories. We have been able to analyze each of the participating platform policies and have established a comparable way to demonstrate a link with the framework.

### Providing contextual insights on data

Our final step has been to provide an understanding around the numbers, explaining overall trends and rationale on changes in the numbers.

# Using the GARM Aggregated Measurement Report

## How should this report be used and how should it not?

Marketers making media decisions today should take responsibility factors into media investment considerations; is the quality and the safety of my reach appropriate for my organization and does it reflect my organization's beliefs and values? This is especially pertinent as it relates to digital media investment. The GARM Aggregated Measurement Report helps create a single resource that collects individual platform transparency reports. While the underlying data is not meant for cross-platform analysis and tabulation, what it can do is provide marketing stakeholders with a single reference in a common language and framework to answer investment considerations related to content safety.

This report should help GARM stakeholders and members do the following:

- Assess safety to inform media selection considerations related to content safety
- Assess progress on safety enforcement
- Assess topical exposure and/or progress
- Determine how to best deploy independent targeting and reporting tools for media campaigns

The report is a useful input tool that creates an even level of understanding on platform safety and advertising. However, this report and the data should not be overused or misused.

- ✗ **Investment Decision Making:** Taken alone, the report is not intended to determine media buying strategies. The report is misused if taken into investment decision making alone (at the expense of more established media reach and cost figures).
- ✗ **Side-by-Side and Direct Comparison:** While the reporting template is harmonized and we have put forth authorized metrics, the underlying policies and timelines between platforms vary. As such it is best to look at the magnitude of the metric and movement, versus direct comparison.
- ✗ **Media Campaign Safety Forecasting and/or Delivery:** The report data is at a global level representing each platform's user base. Media campaigns are typically targeted to users in a geography and focused on a user behavior. As such the generic nature of the data cannot be used to forecast or report on the delivery of a media campaign.

## What is the framework for the report?

GARM's charter celebrates the positive influence of the digital media and advertising industry, but also encourages action to take a more consistent and rigorous approach to curtailing the shadow-side of the industry – specifically the ability of harmful content to reach consumers for brand advertising to appear inadvertently in that environment. With that in mind, we determined there are four core questions for the GARM Aggregated Measurement Report to help the advertising industry answer:

1. How safe is the platform for consumers?
2. How safe is the platform for advertisers?
3. How effective is the platform in policy enforcement?
4. How does the platform perform in correcting mistakes?

In answering these questions, the Measurement and Oversight Working Group within GARM reviewed a series of 80 candidate measures and agreed upon 9 measures that are considered best practices as ‘Authorized Metrics.’ The table below summarizes the recommendations of the working group and secured amongst GARM members:

CORE QUESTION	AUTHORIZED METRIC	DEFINITION + OVERVIEW	RATIONALE
How safe is the platform for consumers?	Prevalence of violating content or Violative View Rate	The percentage of views that contain content that is deemed as violative	Establishes a ratio based on typical user content consumption. Prevalence or Violative View Rate examines views of unsafe/violating content as a proportion of all views.
How safe is the platform for advertisers?	Prevalence of violating content or Advertising Safety Error Rate	The percentage of views that contain content that is deemed as violative The percentage of views of monetized content that contain violative content	Monetization prevalence examines unsafe content viewed as a proportion of monetized content viewed
How effective is the platform in policy enforcement?	Removals of Violating Content + Removal of Violating Accounts Removals of Violating Content expressed by how many times it has been viewed	Pieces of violating content removed Accounts removed due to repeat policy violation Pieces of violating content removed categorized by how many times they were viewed by users	Platform teams spend a considerable amount of time removing violating content and bad actors from their platforms – the magnitude of the efforts should be reported to marketers. It is also important to understand how many times harmful content has been removed.
How does the platform perform at correcting mistakes?	Appeals Reinstatements	Number of pieces of violating content removed that are appealed Number of pieces of violating content removed that are appealed and then reinstated	Platform should be responsive to their users and policy should be consistent with a policy of free and safe speech. For this reason we look at appeals and reinstatement of content removed.

In the event a platform doesn’t have authorized metrics available they are able to provide a measure that is considered to be their next best measure. All of the platforms participating in the GARM Aggregated Measurement Report support the adoption and implementation of the authorized metrics and taking into consideration a development roadmap to fulfill these aspirations. Platforms in GARM will communicate decisions and timelines to adopt Authorized Metrics with the GARM Steer Team via the Measurement and Oversight Working Group.

**How may this report evolve over time?**

Content and advertising safety is a topic that is fluid, and GARM will evolve solutions to address the evolving marketplace and satisfy new needs. As such, the GARM Aggregated Measurement Report will develop undoubtedly over time. We foresee the evolution of the report coming via the following ways:

1. Inclusion of additional GARM platforms in the aggregated measurement report
2. Potential new measures via authorized metrics that help to answer our core questions better
3. Potential specific metrics details at language and/or geographical levels
4. Expansion of GARM content areas to be reported on and tracked

Evolutions to the report will be agreed in GARM via our established governance mechanisms (link here to site content), which will allow for the Measurement and Oversight Working Group to evolve the report for approval by the GARM Steer Team.

We’re excited to launch this report with the partnership and collaboration within GARM, notably with YouTube, Facebook, Instagram, Twitter, TikTok, Snap, Pinterest, Twitch and LinkedIn. For a more detailed overview of how we’ve worked within GARM to create this report, please see the Appendix.

# Executive Summary

Volume 5 of the GARM Aggregated Measurement Report represents the two-year anniversary of our work in the transparency reporting area. We started with the ambitious goals to help the advertising industry better understand the safety enforcement practices of GARM’s platform members. The Aggregated Measurement Report has evolved and improved since its launch in April 2021, thanks to the GARM community, Working Groups, and the Steering Team. Firstly, we’ve compiled data from multiple platforms and focused on the metrics most helpful to advertising stakeholders. Secondly, we’ve seen the Aggregated Measurement Report expand over time – we now have 9 platforms participating, up from 7 when we first started our work. Additionally, we’re seeing more platforms voluntarily adopt the Authorized Metrics the GARM Measurement Working Group in their submissions; when we first launched the Aggregated Measurement Report 44% of the submissions were categorized as using Authorized Metrics, and in this latest one we see 64% of the submissions using Authorized Metrics. We are happy to see more platforms onboard these metrics on behalf of their stakeholders, including advertisers, and we thank the teams who are committing themselves to advancing this work within their companies. We remain thankful to the GARM Working Group who makes this report and the broader improvement on transparency reporting for advertising purposes possible.

## Volume 5 of the report features the following changes:

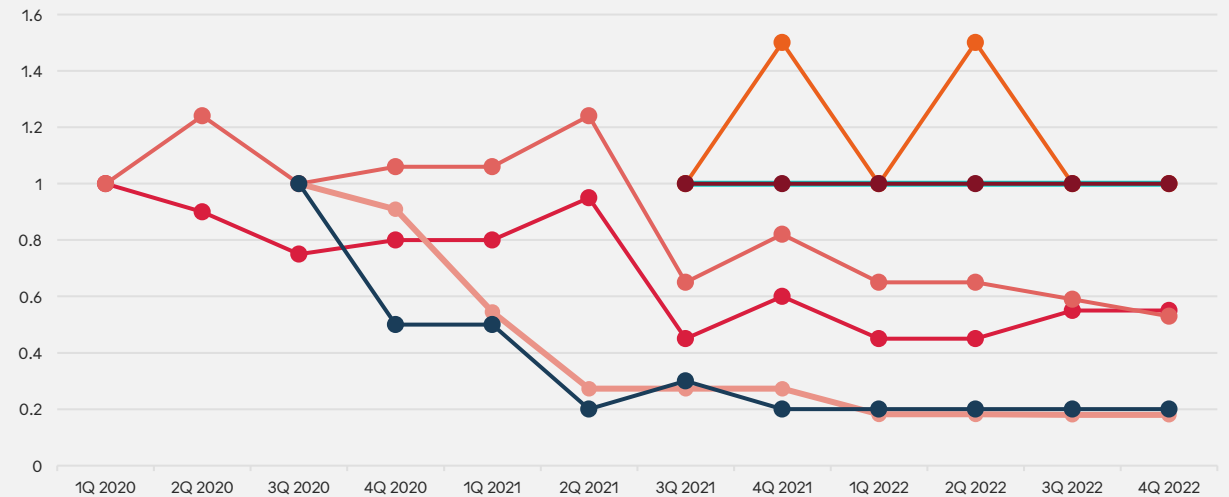
- 01** LinkedIn joins the Aggregated Measurement Report: LinkedIn joined GARM in January 2022. Since then, they’ve taken GARM measurement needs onboard in their broader transparency reporting efforts. We’re pleased to see them join this volume of the Aggregated Measurement Report, and with a submission that has so many authorized metrics.
- 02** Macro insights and data visualization: Another improvement area we are excited to bring forward is our work on industry-level insights, and data visualization. We recognize that the Aggregated Measurement Report is a unique opportunity to look at cross-platform trends in content and enforcement. In our last report we analyzed share of enforcements and we continue to do that in this report. We have also included a visualization to benchmark violative views of content or prevalence over time.

### LEARNING 1:

#### Consumer safety is improving when we look at trends, over time:

When we looked at user safety metrics – typically expressed as Prevalence or Violative View Rate – and trends over time we find modest decreases in the presence of violative content as measured by these metrics. For this visualization, we benchmarked each platform’s consumer safety metric using the earliest Authorized Metric shared by the platform in response to Question 1 of the report. For each successive quarter, we analyzed each platform against that benchmark period via an index. For the purposes of the index, a decrease in Prevalence or Violative View Rate is an improvement in consumer safety where the benchmark is less than 1.0. As a result, we are seeing 4-of-6 measures decrease and 2-of-6 measures stay constant. Notably there has been in large a decrease in Prevalence or Violative View Rate, suggests platforms user safety practices remain active & consistent, as measured in first-party reporting. Our hope is that more platforms report on Advertising Safety Error Rate over time, and we can run similar analyses to focus exclusively on monetization safety.

CHANGES IN USER SAFETY INDEXED TO INITIAL PLATFORM PARTICIPATION IN GARM AGGREGATED MEASUREMENT REPORT



# Executive Summary

## LEARNING 2:

**Share of enforcements analysis shows topics related to individual safety, group hate, and graphic violence emerge as high-volume areas :**

In our last report, we started to look at the number of enforcement actions in specific categories relative to all enforcement actions taken, which we have referred to as ‘Share of Enforcements.’ For the purposes of this analysis share of enforcements takes into account all actions across platforms spanning both content removals and account suspensions. Our previous analysis on share of enforcements demonstrated that two GARM categories dominated enforcement acts-- ”Adult & Explicit Sexual content” and ”Spam or Malware” categories. However, when we move beyond those two common categories and look at share of enforcements in the next 5 largest categories a shift becomes more visible: (1) Crime & Harmful acts to individuals and Society, Human Right Violations; (2) Death, Injury or Military Conflict, and (3) Hate speech & Acts of Aggression. Going back to Volume 4 of the Aggregated Measurement Report, we first saw an increase in Death, Injury & Military Conflict, driven in part by the war in Ukraine.

## TOP SHARE OF ENFORCEMENTS BY GARM CONTENT CATEGORY

- 1 Crime & Harmful Acts to Individuals and Society, Human Right Violations
- 2 Death, Injury or Military Conflict
- 3 Online Piracy
- 4 Hate Speech & Acts of Aggression
- 5 Obscenity And Profanity, Including Language, Gestures, And Explicitly Gory, Graphic or Repulsive Content Intended to Shock and Disgust

## LEARNING 3:

**Enforcement data shows an increase in content removals and a decrease in account removals :**

Analyzing platform enforcement trends showed an increase in content removal by +13%, whereas account removals were down by -16%. The significant increases in content removals are linked to individually-defined and driven platform enforcement efforts across various policies, with a trend towards policies mapped to GARM’s “Crime & Harmful Acts to Individuals & Society, Human Rights Violations” category. Platform shifts in enforcement between content and users are driven by the severity of the platform policy violation and the number of times the user violated platform policies. A table that links GARM Content Categories to each platform’s policies can be found in the Appendix of this report.

## A look forward to Volume 6 and beyond

### Increased focus on Monetization safety trends:

GARM's primary concern is monetization – the placement of ads. In this volume of the report we visualized a trend on user safety. Our hope is that more platforms will focus on reporting metrics specific to monetization safety, and that we will be able to focus trends and visualization on these figures, which are essential to advertising industry stakeholders.

### Increased disclosure on local coverage in global sampling:

We recognize that the core metrics to assess user safety (Prevalence or Violative View Rate) and advertiser safety (Advertising Safety Error Rate or Prevalence) are samples based on global users. In order to drive increased confidence in these methods, GARM will be working to create a disclosure template within the Measurement + Oversight Working Group to improve the transparency in how markets and languages are covered in these sample-based analyses.

### Increased category analysis via Solutions Developers Working Group:

In March 2022, we expanded GARM membership to independent measurement companies that work with digital and social media platforms in the brand safety targeting and reporting capacities. These companies are in a unique position to report on category trends. We will be working with them to add an additional level of analysis on monetization safety and suitability focused on specific content categories and formats to compliment the Aggregated Measurement Report, which at current is focused on platform-specific actions.





## Our Commitment to Responsibility

At YouTube, **Responsibility remains our #1 priority, and we continue to approach this work via our 4 R's of Responsibility strategy: Remove violative content, Raise up authoritative voices, Reduce recommendations of content that brushes right up against our policy line and Reward trusted partners.**

YouTube's commitment to responsibility starts with [Community Guidelines](#) that guide our 'removals' work and set the rules of the road for what we don't allow on our platform. For example, we do not allow pornography, incitement to violence, harassment, hate speech or harmful misinformation. We develop these guidelines in consultation with a wide range of external industry and policy experts and apply them to all types of content on the platform, including videos, comments, links, and thumbnails —regardless of the subject or the creator's background, political viewpoint, position, or affiliation. Our Community Guidelines meet, and often exceed, the GARM Brand Safety Floor ([see page 20](#)).

We enforce these Community Guidelines using a combination of human reviewers and machine learning to remove violative content at scale. In **H2 2022** we were able to detect **>94%** of all violative content on YouTube by automated flagging – with more than two-thirds of flagged content removed with 10 or fewer views. Content flagged by users is only actioned after review by our trained human reviewers to ensure the content does indeed violate our policies and to protect content that has an educational, documentary, scientific, or artistic purpose.

Additionally, we enforce a set of **Ad Friendly Guidelines** policies, which set the standard for which videos are eligible for ads. These guidelines are more restrictive than our Community Guidelines and also adhere to, and often exceed, the [GARM brand safety floor](#). We measure enforcement effectiveness through our Advertiser Safety Error Rate, an MRC-accredited metric also included in this report as a GARM Authorized metric. In 2022, we maintained our commitment to being 99% brand safe as measured by our Advertiser Safety Error Rate. Learn more [here](#).

## Noteworthy ongoing investments

### YouTube Community Guidelines Enforcement Report

Our ongoing quarterly transparency report showcases data demonstrating the impact of our enforcement work and the progress we've made with regards to content that violates our Community Guideline policies; including flagging (human and automated), video, channel, and comment removals, appeals and reinstatements, and highlighted policy verticals. This report first launched in 2018, and like other Google Transparency Reports, the data we share—and the way we share it—evolves over time. Most recently, in Q2 2022 we updated our report to include the number of videos violating our misinformation policies (previously disclosed in the “Spam, Misleading, and Scams” vertical). Our latest Community Guidelines enforcement transparency report can be reviewed [here](#). YouTube highlights data and insights from our last four transparency reports in the GARM Aggregated Measurement Report Volume 5, aggregated as 2H 2022 & 1H 2022.

### European Union Code of Practice on Disinformation Report

As part of our collaboration with the European Union Commission, in early 2023 Google published its first Code of Practice on Disinformation transparency report; highlighting Learn more [here](#), the breadth of YouTube's work across European Economic Area (EEA) Member States to tackle the monetization of disinformation, to provide transparency on political advertising and to work with fact-checking and research communities.

### “Hit Pause” Global Media Literacy Program

In November 2022, YouTube launched a global media literacy campaign to teach viewers critical internet safety skills (e.g. identifying different manipulation tactics used to spread misinformation) via public service announcements (PSAs) and a dedicated [YouTube channel](#).



As of December 2022, the **campaign has been launched in 20 EEA Member States with plans to expand across all EEA Member States in 2023.**

### First Party Brand Suitability Controls for Advertisers

In addition to our brand safety processes, YouTube offers Advertisers access to Suitability controls to achieve unique Advertiser suitability needs.

- In September 2022, we expanded YouTube's 3-Tier inventory mode coverage to YouTube Shorts, automatically extending an Advertiser's selected Inventory Mode (Limited, Standard, Expanded) to Shorts inventory in their YouTube campaigns.
- In October 2022, we streamlined access to our suite of controls by launching the Content Suitability Center within Google Ads, bringing together everything Advertisers need to manage YouTube's Suitability settings for all campaign types on YouTube. Learn more [here](#).

### MRC Content-level Brand Safety Accreditation

In April 2023, the MRC re-accredited YouTube for content-level Brand Safety. YouTube first received this accreditation in 2021, and we take pride in continuing to lead the industry by example in maintaining it annually. The MRC's continued accreditation re-affirms that YouTube in-stream video ads adhere to the industry standards for content level brand safety processes and controls, while validating our Advertiser Safety Error Rate as an accredited metric. Learn more [here](#).



**Report Insights:** In this report, YouTube is proud to answer all four core questions using GARM Authorized Metrics, as we have in previous GARM Aggregated Measurement Report volumes July through December 2022

### July through December 2022

Between July and December 2022, YouTube removed over 11.2 million videos for violating Community Guidelines.

The vast majority (>94%) of these videos were first flagged by machines rather than humans. Over 628k video removals were appealed, and we reinstated <59k of those videos. YouTube terminated over 12.2 million channels for violating our Community Guidelines, the overwhelming majority which were terminated for violating our spam policies. Our VVR ranged from 0.09-0.11% in Q3 and Q4. This means that out of every 10,000 views on YouTube in Q3 and Q4 only 9-11 came from violative content.

YouTube also removed more than **2.6 billion comments**, the majority of which were **spam**; **99%** of removed comments were detected automatically.

Our 2H'22 enforcement efforts were influenced by many factors, including:

- **Policy Updates:** In H2 2022, YouTube continued to update its [Community Guidelines](#) and [Advertiser-friendly Content Guidelines](#). For example, we expanded our Elections misinformation policies within our Community Guidelines, specifically our elections integrity policy, to include the 2022 Brazil Federal elections. We also updated our [Harmful or dangerous acts](#) policies around videos showcasing vigilantism within our Advertiser-friendly content guidelines. A detailed list of YouTube's H2 2022 policy updates can be viewed [here](#).

- **Election Integrity:** During major news moments, YouTube shows content from trusted sources for viewers in prominent news shelves, and the 2022 midterms in the United States, were no exception. For example, in the month of November alone, our Breaking News and Top News shelves surfaced more than 65 million times on the YouTube homepage and at the top of search results in multiple languages, including English and Spanish. In addition to surfacing election-related information panels, shown over 2 billion times, we removed over 10,000 videos related to the midterms for violating our Community Guidelines, including those that violated our election integrity policy. 75% of those removed videos were taken down before they had 100 views.
- **Response to Crises:** YouTube continued to take measures to keep the platform safe during the COVID-19 pandemic and the War in Ukraine. As the Covid-19 situation evolved, YouTube partnered closely with global and local health authorities to ensure policy definition and enforcement was effective in removing violative content where there is a serious risk of egregious harm. In H2 2022, globally, YouTube removed more than 57,000 videos, for violating the vaccine provisions of its COVID-19 Medical Misinformation Policy. These provisions took effect in October 2020. Throughout the war in Ukraine, our teams and systems have continued to restrict and remove harmful content while connecting people to high quality information from authoritative sources. Since Feb. 24, 2022, we've removed more than 9K channels and more than 95k videos related to the ongoing crisis in Ukraine.



### January through June 2022

Between **January and June 2022**, YouTube removed over **8.3 million** videos for violating Community Guidelines. The vast majority (>**92%**) of these videos were first flagged by machines rather than humans. Over **443k** video removals were appealed, and we reinstated **<60k** of those videos. YouTube terminated over **8.3 million** channels for violating our Community Guidelines, the overwhelming majority which were terminated for violating our spam policies. Our VVR ranged from **0.09-0.11% in Q1 and Q2**. This means that out of every 10,000 views on YouTube in **Q1 and Q2 only 9-11** came from violative content.

YouTube also removed more than **1.6 billion comments**, the majority of which were **spam**; **99%** of removed comments were detected automatically.

Our 1H'22 enforcement efforts were influenced by the following factors:

### Tackling Harmful Misinformation

We continue to invest in our work to address harmful misinformation on YouTube. In February 2022, YouTube Senior Vice President, Neal Mohan, shared YouTube's ongoing efforts to combat misinformation challenges in a [blog post](#) published on the YouTube Official blog. These efforts include strengthening our systems by training them on new data to catch misinformation before it goes viral, and connecting viewers to authoritative videos in search results and recommendations.

Beyond growing our teams with even more people who understand the regional nuances entwined with misinformation, we're exploring further investments in partnerships with experts and non-governmental organizations around the world. We'll continue to rigorously enforce our policies through a combination of human review and machine learning technology.

In order to provide transparency into how we handle misinformation on our platform, we updated our [Community Guidelines Enforcement Report](#) to include the number of videos removed for violating our [misinformation policies](#). For example, this includes medical and general misinformation. We previously disclosed these removals under the "Spam, Misleading and Scams" vertical in the Community Guidelines Enforcement Report.

In Q2 2022, we removed more than 122,000 videos for violating these policies, which includes the removal of 35K videos for violating the vaccine provisions of our COVID-19 misinformation policy that took effect in October 2020. In GARM's Aggregated Measurement Report Volume 5, enforcement of our misinformation policies is represented as part of YouTube's "Spam, deceptive practices, scams and misinformation" category.

### Responding to the War in Ukraine

Throughout the war in Ukraine, our teams and systems have continued to restrict and remove harmful content while connecting people to high quality information from authoritative sources:

- **Policy & Enforcement:** We remove content about the war in Ukraine that violates our Community Guidelines—including content that violates our major violent events policy, which prohibits content denying, minimizing or trivializing that a well-documented, violent event took place and which we [expanded](#) to include Russia's invasion in Ukraine. We've removed more than 9,000 channels and more than 76,000 videos related to the war for violating our Community Guidelines and Terms of Service, and restricted channels associated with Russian state-funded news channels globally, resulting in more than 750 channels and more than 4 million videos blocked.
- **Raising Authoritative Sources:** We're connecting viewers to high-quality information about the war in Ukraine, by raising videos from authoritative sources in search results and recommendations. Our breaking news and top news shelves on our homepage have received more than 75 million views in Ukraine.

Our teams continue to closely monitor the war and are ready to take further action. More information on our efforts to help Ukraine can be found on The Google Keyword [Blog](#). **In GARM Aggregated Measurement Report Volume 5, our ongoing Ukraine enforcement efforts are reflected in our broader enforcement of policies such as "Harmful & Dangerous", "Violent or Graphic", and "Spam, deceptive practices, scams, and misinformation" categories.**



### Methodology for Metrics

In this resource, we've offered various metrics to answer the four key questions we know marketers are asking about platform responsibility. Below is a summary of how we define and calculate each metric:

**Violative View Rate:** The Violative View Rate (VVR) represents the percentage of views on YouTube that come from content that violates our Community Guidelines policies.

**Removed Videos:** YouTube relies on teams around the world to review flagged videos and remove content that violates our Community Guidelines. This exhibit shows the number of videos removed by YouTube for violating its Community Guidelines per quarter.

**Removed Videos by Views:** This chart shows the percentage of video removals that occurred before they received any views versus those that occurred after receiving some views.

**Removed Videos by Views (as first detected by machines):** Automated flagging enables us to act more quickly and accurately to enforce our policies. This chart shows the percentage of video removals, that were first detected by machines, that occurred before they received any views versus those that occurred after receiving some views.

**Advertiser Safety Error Rate:** This metric indicates how often unsafe content is incorrectly monetized and is calculated as follows:

- Brand safety error rate = # of impressions on unsafe content / # total impressions
- We take 1000 impression-weighted random samples a day (for 5 days a week) from across all ad impressions on YouTube. We then calculate the brand safety error rate as a 60-day average across all 60,000 impressions
- Each impression is associated with one video, which is human reviewed by trained raters and given a Brand Safety decision.

YouTube's Advertiser Safety Error Rate was included in the MRC Content Level Brand Safety Controls Audit, and in YouTube's annual MRC recertification for May 2022; specific to ads sold through Google Ads, Display & Video 360 (DV360) and YouTube Reserve, including in-stream ads and excluding video discovery, masthead, YouTube Kids and livestream.

**Removed Comments:** Using a combination of people and technology, we remove comments that violate our Community Guidelines. We also filter comments which we have high confidence are spam into a 'Likely spam' folder that creators can review and approve if they choose.

This exhibit shows the volume of comments removed by YouTube for violating our Community Guidelines and filtered as likely spam which creators did not approve. The data does not include comments removed when YouTube disables the comment section on a video.

It also does not include comments taken down when a video itself is removed (individually or through a channel-level suspension), when a commenter's account is terminated, or when a user chooses to remove certain comments or hold them for review.

**Removed Channels:** A YouTube channel is terminated if it accrues three Community Guidelines strikes in 90 days, has a single case of severe abuse (such as predatory behavior), or is determined to be wholly dedicated to violating our guidelines (as is often the case with spam accounts). When a channel is terminated, all of its videos are removed.

This exhibit shows the number of channels removed by YouTube for violating its Community Guidelines per quarter."



**Videos appealed:** If a creator chooses to submit an appeal, it goes to human review, and the decision is either upheld or reversed.

This exhibit shows the number of appeals YouTube received for videos removed due to a Community Guidelines violation per quarter. Creators have 30 days to submit an appeal after the video's removal, so this number also includes appeals for videos removed during one quarter but appealed in the following quarter.

**Appealed videos reinstated:** If a creator chooses to submit an appeal, it goes to human review, and the decision is either upheld or reversed. The appeal request is reviewed by a senior reviewer who did not make the original decision to remove the video. The creator receives a follow up email with the result.

This exhibit shows the number of videos YouTube reinstated due to an appeal after being removed for a Community Guidelines violation per quarter. Note that a reinstatement counted here may be in response to an appeal or video removal that occurred in a previous quarter.



## Question 1: How safe is the platform for consumers?

### Authorized Metric: Violative View Rate

Violative View Rate is an estimate of the proportion of video views that violate YouTube's Community Guidelines in a given quarter (excluding spam)

GARM Metric	Latest Period		Previous Period	
	Q3 2022	Q4 2022	Q1 2022	Q2 2022
Violative View Rate	0.10%-0.11%	0.09%-0.11%	0.09-0.11%	0.09-0.11%

YouTube consistently makes improvements to our methodology to more accurately calculate VVR. This exhibit reflects the most current methodology used to calculate VVR as of the time period reported. Secondly, if our Community Guidelines expand to include a new type of violative content in the future, VVR will increase to reflect this expanded scope, as our systems learn to detect this new type of content



## Question 2: How safe is the platform for advertisers?

**Authorized Metric:** Advertising Safety Error Rate



Advertiser Safety Error Rate is the percentage of total impressions on content that is violative of our monetization policies – which align with the GARM industry standards – for in-stream content

GARM Metric	Latest Period		Previous Period	
	Q3 2022	Q4 2022	Q1 2022	Q2 2022
Advertising Safety Error Rate	<1%	<1%	<1%	<1%





### Question 3a: How Effective is the Platform in Enforcing Safety Policy?

**Authorized Metric:** Content Actioned, Actors Actioned, Comments Actioned, Removal of Videos by view

Violating content acted upon and removed by YouTube and the percentage of removed videos by views and the percentage of views as first detected by machines

#### YouTube Community Guidelines

- Guidelines governs content that can live on YouTube.
- Enforcement of these guidelines is reflected in our quarterly [Community Guidelines Enforcement Report](#)

YouTube Policy	Content Actioned <sup>1</sup>		Actors Actioned <sup>2</sup>		Comments Actioned	
	Latest Period Q3 & Q4 2022	Previous Period Q1 & Q2 2022	Latest Period Q3 & Q4 2022	Previous Period Q1 & Q2 2022	Latest Period Q3 & Q4 2022	Previous Period Q1 & Q2 2022
Nudity or sexual	1,452,591	1,320,730	323,343	335,801	3,086,895	2,365,605
Child safety	3,941,120	2,351,206	197,876	160,614	200,961,384	194,720,777
Harmful or dangerous	1,608,999	1,015,223	38,036	61,847	1,829,276	118,713
Promotion of violence and violent extremism	141,550	133,711	20,001	19,890	1,140,429	446,506
Harassment and cyberbullying	1,033,799	922,039	98,286	119,999	165,577,995	255,959,014
Violent or graphic	1,962,864	1,724,913	87,281	18,021	1,772,831	44,356
Spam, deceptive practices, scams and misinformation	701,188	641,280	11,328,249	7,538,816	2,218,538,886	1,121,879,591
Hateful or abusive	373,825	241,635	86,390	75,559	77,022,492	122,379,812
Impersonation	n/a	n/a	67,989	47,741	n/a	n/a
Other	51,033	28,880	22,818	19,570	348	200

<sup>1</sup> Content Actioned for YouTube is Videos Removed

<sup>2</sup> Actors Actioned for YouTube is Channels Removed



### Question 3b: How Effective is the Platform in Enforcing Safety Policy?

**Authorized Metric:** Content Actioned, Actors Actioned, Comments Actioned, Removal of Videos by view

Violating content acted upon and removed by YouTube and the percentage of removed videos by views and the percentage of views as first detected by machines

GARM Metric	Latest Period	Previous Period
	Q3 & Q4 2022	Q1 & Q2 2022
Total Video Removals	11,266,969	8,379,617
Removed videos by views: 0 views	36.3%	32.8%
Removed videos by views: 1-10	32.0%	35.1%
Removed videos by views: 10+	31.7%	32.1%





#### Question 4: How does the platform perform at correcting mistakes?

**Authorized Metric:** Appeals, Reinstatements

YouTube measures correction of mistakes by the number of video appeals and number of video reinstatements

GARM Metric	Latest Period	Previous Period
	Q3 & Q4 2022	Q1 & Q2 2022
Content Appealed: Videos	628,665	443,507
Content Reinstated: Video	58,417	59,475



# Mapping GARM Categories and Monetization to YouTube Community Policy-level Reporting

In the YouTube Community Guidelines Enforcement Report, Video, Comment and Channel removals are broken down by Community Guideline removal reason. In the table below, we have mapped each of these removal reasons to the most complementary GARM Brand Safety Floor category as a reference point for you. Remember, though: **our Community Guidelines set the rules of the road for what we allow on our platform. The GARM Brand Safety Floor – to which our Ad Friendly Guidelines are aligned – set the standard for which videos are eligible for ads on YouTube.** Our Community Guidelines Enforcement Report offers data on the enforcement of our Community Guidelines, not our Ad Friendly Guidelines. We offer this table to help you understand how our Community Guidelines definitions compare with GARM's definitions of brand unsafe content.

<p><b>GARM Brand Safety Floor Category + Definition</b></p> <ul style="list-style-type: none"> <li>• Defines content that can monetize.</li> <li>• Aligned with <a href="#">YouTube's Ad Friendly Guidelines</a>, a higher bar than Community Guidelines.</li> </ul>	<p><b>Relevant <a href="#">YouTube Community Guidelines</a></b></p> <ul style="list-style-type: none"> <li>• Governs content that can live on YouTube.</li> <li>• Our <a href="#">Community Guidelines Enforcement Report</a> measures our enforcement of these guidelines.</li> </ul>
<p><b>Adult &amp; Explicit Sexual Content</b></p> <ul style="list-style-type: none"> <li>• Illegal sale, distribution, and consumption of child pornography</li> <li>• Explicit or gratuitous depiction of sexual acts, and/or display of genitals, real or animated</li> </ul>	<p><b>Nudity and sexual Content</b> Explicit content meant to be sexually gratifying is not allowed on YouTube. Posting pornography may result in content removal or channel termination. Videos containing fetish content will be removed or age-restricted. In most cases, violent, graphic, or humiliating fetishes are not allowed on YouTube.</p> <p><b>Child safety</b> YouTube doesn't allow content that endangers the emotional and physical well-being of minors. A minor is defined as someone under the legal age of majority -- usually anyone younger than 18 years old in most countries/regions.</p>
<p><b>Arms &amp; Ammunition</b></p> <ul style="list-style-type: none"> <li>• Promotion and advocacy of Sales of illegal arms, rifles, and handguns</li> <li>• Instructive content on how to obtain, make, distribute, or use illegal arms</li> <li>• Glamorization of illegal arms for the purpose of harm to others</li> <li>• Use of illegal arms in unregulated environments</li> </ul>	<p><b>Firearms</b> Content intended to sell firearms, instruct viewers on how to make firearms, ammunition, and certain accessories, or instruct viewers on how to install those accessories is not allowed on YouTube. YouTube shouldn't be used as a platform to sell firearms or accessories noted below. YouTube also doesn't allow live streams that show someone holding, handling, or transporting a firearm.</p>
<p><b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b></p> <ul style="list-style-type: none"> <li>• Graphic promotion, advocacy, and depiction of willful harm and actual unlawful criminal activity - Explicit violations/demeaning offenses of Human Rights (e.g. human trafficking, slavery, self-harm, animal cruelty etc.)</li> <li>• Harassment of bullying of individuals and groups</li> </ul>	<p><b>Harmful or dangerous Content</b> YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death.</p> <p><b>Hate speech</b> Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status</p> <p><b>Harassment and cyberbullying</b> Content that threatens individuals is not allowed on YouTube. We also don't allow content that targets an individual with prolonged or malicious insults based on intrinsic attributes. These attributes include their protected groups or physical traits.</p>
<p><b>Death, Injury or Military Conflict</b></p> <ul style="list-style-type: none"> <li>• Promotion, incitement or advocacy of violence, death or injury</li> <li>• Murder or willful bodily harm to others</li> <li>• Graphic depictions of willful harm to others</li> <li>• Incendiary content provoking, enticing, or evoking military aggression</li> <li>• Live action footage/photos of military actions &amp; genocide or other war crimes</li> </ul>	<p><b>Violent or graphic content</b> Violent or gory content intended to shock or disgust viewers, or content encouraging others to commit violent acts are not allowed on YouTube.</p> <p><b>Harmful or dangerous content</b> YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death.</p> <p><b>Suicide &amp; self-injury</b> We do not allow content on YouTube that promotes suicide, self-harm, or is intended to shock or disgust users.</p>





# Mapping GARM Categories and Monetization to YouTube Community Policy-level Reporting

<b>Online piracy</b> <ul style="list-style-type: none"> <li>• Pirating, Copyright infringement, &amp; Counterfeiting</li> </ul>	<b>Fake engagement</b> YouTube doesn't allow anything that artificially increases the number of views, likes, comments, or other metric either through the use of automatic systems or by serving up videos to unsuspecting viewers. Additionally, content that solely exists to incentivize viewers for engagement (views, likes, comments, etc.) is prohibited.
	<b>Impersonation</b> Content intended to impersonate a person or channel is not allowed on YouTube. YouTube also enforces trademark holder rights. When a channel, or content in the channel, causes confusion about the source of goods and services advertised, it may not be allowed.
	<b>Sale of illegal or regulated goods or services</b> Content intended to sell certain regulated goods and services is not allowed on YouTube. Such as: Counterfeit documents or currency
	<b>YouTube's Terms of Service</b> Also covered in YouTube's Terms of Service
<b>Hate speech &amp; acts of aggression</b> <ul style="list-style-type: none"> <li>• Behavior or content that incites hatred, promotes violence, vilifies, or dehumanizes groups or individuals based on race, ethnicity, gender, sexual orientation, gender identity, age, ability, nationality, religion, caste, victims and survivors of violent acts and their kin, immigration status, or serious disease sufferers.</li> </ul>	<b>Hate speech</b> Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability (including chronic or lifelong diseases), Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b> <ul style="list-style-type: none"> <li>• Excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult.</li> </ul>	<b>Violent or graphic content</b> Violent or gory content intended to shock or disgust viewers, or content encouraging others to commit violent acts are not allowed on YouTube.
	<b>Age restriction</b> Sometimes content doesn't violate our policies, but it may not be appropriate for viewers under 18. In these cases, we may place an age-restriction on the video. This policy applies to videos, video descriptions, custom thumbnails, live streams, and any other YouTube product or feature. For example, this can include content with vulgar language.
<b>Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol</b> <ul style="list-style-type: none"> <li>• Promotion or sale of illegal drug use - including abuse of prescription drugs. Federal jurisdiction applies, but allowable where legal local jurisdiction can be effectively managed</li> <li>• Promotion and advocacy of Tobacco and e-cigarette (Vaping) &amp; Alcohol use to minors</li> </ul>	<b>Sale of illegal or regulated goods or services</b> Content intended to sell certain regulated goods and services is not allowed on YouTube. Such as: controlled narcotics and other drugs, nicotine, including vaping products, pharmaceuticals without a prescription, unlicensed medical services
	<b>Harmful or dangerous content</b> YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death.
<b>Spam or Harmful Content</b> <ul style="list-style-type: none"> <li>• Malware/Phishing</li> </ul>	<b>Spam deceptive practices, scams and misinformation</b> YouTube doesn't allow spam, scams, or other deceptive practices that take advantage of the YouTube community. We also don't allow content where the main purpose is to trick others into leaving YouTube for another site. Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes certain types of misinformation that can cause real-world harm, like promoting harmful remedies or treatments, certain types of technically manipulated content, or content interfering with democratic processes.
<b>Misinformation</b> <ul style="list-style-type: none"> <li>• The presence of verifiably false or willfully misleading content that is directly connected to user or societal harm</li> </ul>	
<b>Terrorism</b> <ul style="list-style-type: none"> <li>• Promotion and advocacy of graphic terrorist activity involving defamation, physical and/or emotional harm of individuals, communities, and society</li> </ul>	<b>Violent criminal organizations</b> Content intended to praise, promote, or aid violent criminal organizations is not allowed on YouTube. These organizations are not allowed to use YouTube for any purpose, including recruitment.
<b>Debated Sensitive Social Issue</b> <ul style="list-style-type: none"> <li>• Insensitive, irresponsible and harmful treatment of debated social issues and related acts that demean a particular group or incite great conflict</li> </ul>	
<b>Other</b>	<b>Other</b> Any categories not specifically accounted for in the above mentioned categories. For example, Other would be used to capture a channel that was removed for violating multiple policies.





### Our Commitment to Responsibility and Transparency on Facebook and Instagram

We want Facebook and Instagram to be places where people can express themselves. To make this possible, we must protect our community’s safety, privacy, dignity and authenticity. This is why we have [Community Standards on Facebook](#) and [Community Guidelines on Instagram](#) that define what content is and is not allowed. We take action on content that goes against these policies, and we invest in technology, processes and people to help us act so any violations impact as few people as possible. These policies either meet or, in many cases, exceed the [GARM Brand Safety Floor](#). Facebook and Instagram share content policies, which means that if content is considered violating on one platform, it is also considered violating on the other. Our Community Standards and Community Guidelines apply to all content on our platforms (such as posts, photos, videos or comments). We believe that it’s important that we show the areas where we need to continue to make progress, which is why we were one of the first platforms in 2018 to [begin publishing metrics](#) at a policy level detailing the prevalence of violating content we missed, the content actioned (and the percentage of that we found proactively), and the content appealed and restored. Our Q1 2023 report was our 17th report, and we [celebrated 5 years](#) of releasing this report in May 2023.

We scale our enforcement to review millions of pieces of content across the world every day and use our technology to help detect and [prioritize content that needs review](#). We have around 40,000 people who continue to focus on safety and security efforts, we continue to have teams around the world review content in over 70 languages — most of which are reviewed 24/7. We continue to build technologies that help us identify harmful content faster, across languages and different content types. Our continued focus on AI research helps our technology scale quickly to keep our platforms safe, and our multi-year investments have helped us to build teams that develop policies, improve our technologies, and respond to real-world developments. These tactics have enabled us to cut hate speech prevalence by more than 80% since we began reporting it on Facebook, and we’re using these same tactics across policy areas like violence and incitement and bullying and harassment. (For the definition of prevalence and how it is calculated see our metrics definitions at the end.) To better address hate speech, bullying and harassment and violence and incitement — all of which require understanding of language, nuance and cultural norms — we deployed a [cross-problem AI system](#) to consolidate learnings for all three to better address each violation area. We reduce prevalence of violating content in a number of ways, including developments like these in detection and enforcement and [reducing problematic content in Feed](#).

### Actions we have Taken

We are always refining our policies and enforcement so that we’re both supporting people’s ability to express themselves and protecting safety across our platforms. We found that using [warning screens](#) to discourage hate speech or bullying and harassment content prevented some of this content — which could have violated our Community Standards — from being posted. We improved the accuracy of our [AI technology](#), which resulted in a decrease in hate speech-related content in Q3 2022. We did this by leveraging data from past user appeals to identify posts that could have been removed by mistake without appropriate cultural context. Similarly, our actions against content that incites violence decreased in Q3 2022 after our improved AI technology was able to better recognize language and emojis used in jest between friends. Our steady improvements can be attributed to a holistic approach that includes:

- Development of our policies and ongoing refinement to ensure they best serve the needs of the people using our technologies
- The algorithms and artificial intelligence that help us enforce at scale
- An approach to product design that focuses on safety and integrity
- Making it easy for users to both report and appeal content decisions to help us continue to improve

### Youth Safety and Well-Being

Ensuring that young people have positive and age-appropriate experiences is a responsibility we take seriously. We want to strike the right balance of giving young people freedom on Instagram and Facebook, while also keeping them safe. We recognize that younger users require additional safeguards for their safety, privacy and well-being and our approach towards this is expansive.



We [ground our approach](#) in research, direct feedback from parents, teens, experts, UN children’s rights principles and global regulation. We’ve developed more than 30 tools to support teens and their families. To keep young people safe::

- We set teens’ accounts to private when they join Instagram.
- We limit the amount of potentially sensitive content they can see in Explore, Search and Reels.
- We don’t allow content that promotes suicide, self-harm or eating disorders, and take action on 98% of that type of content we identify before it’s reported to us.

To help parents and teens navigate social media together:

- We have parental tools that let parents and guardians see who their teen reports or blocks, and set “blocking hours” for when they can use our platforms.
- We launched Family Center with expert resources on how to have smart dialogues with teens about online habits.

To give people ways to manage their time so it’s intentional and meaningful:

- We give people the option to turn on ‘Take a Break’ on Instagram to remind them to take regular breaks - and we send teens notifications to do so.
- We notify teens that it might be time to look at something different if they’ve been scrolling on the same topic for a while on Instagram .

### Misinformation

To address misinformation we have built the largest [global fact-checking network](#) of any platform, with more than 90 fact-checking partners around the world who review and rate viral misinformation. In Q2 2022, we displayed warnings on over 200 million distinct pieces of content on Facebook (including reshares) globally based on over 130,000 debunking articles written by our fact-checking partners. In the US, we partner with 10 fact-checking organizations, five of which cover content in Spanish.

### Oversight Board

Prompted by feedback from the Oversight Board, we [shared more details](#) about steps we’ve taken to update Facebook’s penalty system to make it more fair and effective. Under the new system, we will focus on helping people understand why we have removed their content, which is shown to be more effective at preventing re-offending. We are still removing violating content just as we did before, but now we’re also giving people the chance to change their behavior while still applying stronger penalties to more severe violations: posting content that includes terrorism, child exploitation, human trafficking, serious suicide promotion, sexual exploitation, the sale of non-medical drugs, or the promotion of dangerous individuals and organizations. This leads to faster and more impactful actions for those that continuously violate our policies. These changes follow feedback from our community — including our [civil rights auditors](#), the Oversight Board and independent experts — who noted that our current systems needed better balance between punishing and encouraging remediation through education.

### Monetization Policies

On our platforms there are areas where content is eligible to be monetized, so we have [Partner Monetization Policies](#) and [Content Monetization Policies](#) that determine what content can be monetized, and which partners are eligible to earn revenue from ad placement - so even though the content may be allowed on our platforms through our Community Standards and Guidelines, we may determine based on our Content Monetization Policies that it cannot be monetized. These policies are aligned to the [GARM Suitability Framework](#). We also have [Advertising Standards](#) in our [Transparency Center](#) that provide policy detail and guidance on the types of ad content we allow, and the types of ad content we prohibit. Our Advertising Standards also provide guidance on advertiser behavior that may result in advertising restrictions being placed on a business account or its assets (an ad account, Page or user account).

A full view of our efforts on Safety and Integrity are captured [at this timeline](#). While we have good progress to highlight, there is always room for improvement.



### Key trends in Q4 2022 data

In the Q4 2022, we continued to make progress on removing content that violates our Community Standards. Prevalence of harmful content on Facebook and Instagram remained relatively consistent or decreased from Q3 2022 to Q4 2022 across most of our policy areas, meaning the vast majority of content that users encounter does not violate our standards. We updated our [cross-problem AI system](#), combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.

### Independent verification and working with external experts

Our transparency reports allow the public to hold us accountable and help us improve how we talk about our work. We are also committed to undertaking and releasing independent, third-party assessments for our processes, policies and metrics.

We [released](#) an EY assessment of our Community Standards Enforcement Report, which concluded that the calculation of the metrics in the report were fairly stated, and that our internal controls are suitably designed and operating effectively. In November 2022, we [announced](#) that we received accreditation from the MRC for content-level Brand Safety on Facebook. We will be expanding the scope of MRC's audit to include additional advertiser-facing controls (like the inventory filters for Facebook and Instagram Feeds) as we make them more widely available.

Meta assumed the chair of the [Global Internet Forum to Counter Terrorism](#) (GIFCT)'s Operating Board in January 2023. GIFCT is an NGO that brings together technology companies to tackle terrorist content online through research, technical collaboration and knowledge sharing. Meta is a founding member of GIFCT, which was established in 2017 and brings together member companies, governments and civil society organizations to tackle terrorist and violent extremist content online. Meta is also made available a free open source software tool it developed that will help platforms identify copies of images or videos and take action against them en masse. We hope the tool — called [Hasher-Matcher-Actioner \(HMA\)](#) — will be adopted by a range of companies to help them stop the spread of terrorist content on their platforms, and will be especially useful for smaller companies who don't have the same resources as bigger ones. HMA builds on Meta's [previous open source image and video matching software](#), and can be used for any type of violating content.

We [also announced](#) that Instagram and Facebook are founding members of [Take It Down](#) — a new platform by NCMEC to help prevent young people's intimate images from being posted online in the future. Take It Down lets young people take back control of their intimate images. Built in a way that respects young peoples' privacy and data security, Take It Down allows people to only submit a hash — rather than the intimate image or video itself — to NCMEC. Hashing turns images or videos into a coded form that can no longer be viewed, producing hashes that are secure digital fingerprints. Take It Down was designed with Meta's financial support. We are working with NCMEC to promote Take It Down across our platforms, in addition to integrating it into Facebook and Instagram so people can easily access it when reporting potentially violating content. Take It Down builds off of the success of platforms like [StopNCII](#), a platform [we launched](#) in 2021 with South West Grid for Learning (SWGfL) and more than 70 NGOs worldwide, which helps adults stop the spread of their intimate images online, a practice commonly referred to as "revenge porn."

We will continue in our mission to lead the industry in transparency efforts and to provide independent review across both our transparency reporting and our advertiser controls. We collaborate with the industry to align on industry standards around safety and suitability, and we support independent oversight to hold us accountable.





## Meta Brand Safety & Suitability Progress

### First Party Inventory Filters for Facebook and Instagram Controls for Feed

In March, we [announced](#) that Meta's new inventory filters for Facebook and Instagram Feeds are rolling out to advertisers in English and Spanish-speaking markets. AI is one of the driving forces behind these industry-leading solutions. We've spent many years working closely with partners in the industry, including the Global Alliance for Responsible Media (GARM). We've developed controls that align with GARM's Suitability Framework, which defines high, medium and low risk content. In July, we expanded these controls to support additional languages and made them available in more countries. We've also started to test the expansion of Inventory Filter to support additional placements, like Reels, and we'll continue to expand to other surfaces across Facebook and Instagram as we learn more about advertiser preferences to improve and enhance this technology.

### Third Party Brand Suitability Verification in Feed

In March 2023, we also [announced](#) an independent AI-powered solution with our Meta business partner, Zefr, to report the suitability of content adjacent to ads on Facebook Feed. In early testing, we found through third party verification with Zefr, that less than one percent of content on Facebook Feed falls into the high risk GARM suitability category. In July 2023, we expanded this integration to also support Instagram Feed in addition to Facebook Feed. Zefr's AI product assesses video, image, text and audio to label Feed content aligned with the GARM suitability framework. The solution allows advertisers to measure, verify and understand the suitability of content near their ads to help them make informed decisions in order to reach their marketing goals. Meta will be rolling out this verification and measurement solution to additional badged Meta Business Partners this year.

We will be expanding the scope of MRC's audit to include additional advertiser-facing controls (like the inventory filters for Facebook and Instagram Feeds) as we make them more widely available.

## Transparency Reporting and Methodology

As a single destination for our integrity and transparency efforts, last year we launched the [Transparency Center](#). It includes information on:

- [Our policies](#) and how they are developed and updated
- [Our approach to enforcing these content policies](#), using reviewers and technology
- Deep dives on how we work to [safeguard elections and combat misinformation](#)
- [Reports sharing data on our efforts](#) (including the Community Standards Enforcement Report)

Our [Community Standards Enforcement Report](#) metrics definitions:

- Prevalence: How prevalent were violation views on our services?
  - Shows the potential of violating content actually being seen
  - Calculated as the estimated number of views that showed violating content, divided by the estimated number of total content views on Facebook or Instagram
  - We use stratified and random sampling to find the estimated number of views of how much violating content is on our platforms. The sampling is done by manual (human) review.
  - Both sampling types have a 95% confidence window
  - Where the violation type is very infrequent, we use an upper-bound prevalence number (e.g., under 0.006%) rather than a range of values (e.g., 0.05%-0.06%)
  - To generate a representative measurement of global prevalence, we sample and label content in the multiple languages for Facebook and Instagram and are confident this approach provides a representative global estimate.



- Content Actioned: How much content did we take action on?
  - Shows the number of pieces of content (such as posts, photos, videos or comments) we took action on. Actions may include removing content, covering content with a warning screen or disabling an account.
  - Shows the scale of our enforcement activity
  - Content actioned doesn't indicate how much of that violating content actually affected users (that information is captured in prevalence)
- Proactive Rate: How much violating content did we find before users reported it?
  - It shows of the content we took action on, how much we found before it was reported to us
  - A measure of how effective we are at detecting violations and should be viewed in tandem with content actioned
  - When this number is low, it means that our AI is still in the early stages of development. When it is high, it shows that we are doing a better job of finding this content before it was reported.
- Appealed Content: How much of the content we actioned did people appeal?
  - The number of pieces of content (such as posts, photos, videos or comments) that people appeal after we take action on it for going against our policies
  - Numbers can't be compared directly between content actioned or to content restored for the same quarter. Some restored content may have been appealed in the previous quarter, and some appealed content may be restored in the next quarter.
- Restored Content: How much content did we restore after taking action on it, before or after an appeal?
  - The number of pieces of content (such as posts, photos, videos or comments) we restored after we originally took action on them
  - We report content that we restored in response to appeals, as well as content we restored that wasn't directly appealed
  - By "restore," we mean returning content to Facebook that we previously removed or removing a cover from content that we previously covered with a warning

Prevalence is the main metric we hold our teams accountable to as it shows how often people see harmful content on our platform. We report on how much harmful content is seen rather than how much is posted, because we want to determine how much that harmful content actually affected users on our platforms. We evaluate the effectiveness of our enforcement by trying to keep the prevalence of violating content on our platform as low as possible, while minimizing mistakes in the content that we remove. We were the first in the industry to release prevalence metrics, and are pleased to see that several other companies have adopted it as well (sometimes call "violative view rate").

For more details about our processes, methodologies and how we arrived at the numbers, visit our [Transparency Center](#).



## Mapping of GARM Brand Safety Floor to Facebook Community Standards

GARM/4As Category	Facebook Policy		
Adult and Explicit Sexual Content	<a href="#">Adult Nudity and Sexual Activity</a> , <a href="#">Child Sexual Exploitation</a> , <a href="#">Abuse and Nudity</a> , <a href="#">Sexual Solicitation</a>		
Arms and Ammunition	<a href="#">Violence and Incitement</a> , <a href="#">Coordinating Harm and Promoting Crime</a> , <a href="#">Restricted Goods and Services</a>		
Crime and Harmful Acts to Individuals and Society and Human Right Violations	<a href="#">Adult Nudity and Sexual Activity</a> , <a href="#">Violence and Incitement</a> , <a href="#">Bullying and Harassment</a> , <a href="#">Violent and Graphic Content</a> , <a href="#">Child Sexual Exploitation</a> , <a href="#">Abuse and Nudity</a> , <a href="#">Suicide and Self-Injury</a> , <a href="#">Human Exploitation</a> , <a href="#">Dangerous Individuals and Organizations</a> , <a href="#">Coordinating Harm and Promoting Crime</a> , <a href="#">Restricted Goods and Services</a> , <a href="#">Fraud and Deception</a>		
Death, Injury or Military Conflict	<a href="#">Violence and Incitement</a> , <a href="#">Violent and Graphic Content</a> <a href="#">Suicide and Self-Injury</a>		
Online Piracy	<a href="#">Intellectual Property</a> , <a href="#">Fraud and Deception</a>		
Hate Speech and Acts of Aggression	<a href="#">Hate Speech</a> , <a href="#">Bullying and Harassment</a> , <a href="#">Dangerous Individuals and Organizations</a>		
Obscenity and Profanity, including language, gestures and explicitly gory, graphic or repulsive content intended to shock and disgust	<a href="#">Hate Speech</a> , <a href="#">Bullying and Harassment</a>		
Illegal Drugs/Tobacco/E-cigarettes/ Vaping/Alcohol	<a href="#">Restricted Goods and Services</a>		
Spam or Harmful Content	<a href="#">Cybersecurity</a> , <a href="#">Spam</a>		
Terrorism	<a href="#">Dangerous Individuals and Organizations</a>		
Debated Sensitive Social Issues	<a href="#">Hate Speech</a> , <a href="#">Bullying and Harassment</a>		
Misinformation	<a href="#">Misinformation</a>		
<b>Additional policies not covered</b>	<b>Facebook Policy</b>		
Floor focuses online and not on offline/real-world fraud	<a href="#">Fraud and Deception</a>		
Floor does not include census and voter interference/fraud	<a href="#">Coordinating Harm and Promoting Crime</a>		
Floor does not include coverage for creepshots	<a href="#">Adult Sexual Exploitation</a>		
Other Facebook Policies Floor does not address	<a href="#">Privacy Violations</a> <a href="#">Account Integrity and Authentic Integrity</a>	<a href="#">Inauthentic Behavior</a> <a href="#">Memorialization</a>	<a href="#">User Requests</a> <a href="#">Additional Protections for Minors</a>



**Question 1:** How safe is the platform for consumers?

**Question 2:** How safe is the platform for advertisers?

**Authorized Metric:** Prevalence

Prevalence shows the potential of violating content actually being seen. Prevalence considers all the views of content on Facebook and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2022	Q3 2022	Q2 2022	Q1 2022	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	0.06%	0.05%	0.04%	0.04%	<p>Adult Nudity and Sexual Activity: Prevalence increased in Q4 2022 due to bugs within our systems that have now been mitigated in addition to adjustments made to our proactive detection technology.</p> <p>Child Endangerment: Nudity and Physical Abuse and Child Endangerment: Sexual Exploitation. We cannot estimate prevalence for these right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.</p>
	Child Endangerment: Nudity and Physical Abuse	N/A	N/A	N/A	N/A	
	Child Endangerment: Sexual Exploitation	N/A	N/A	N/A	N/A	
Arms & Ammunition	Restricted Goods and Services: Firearms	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	<p>Violence and Incitement: Prevalence decreased in Q4 2022 due to improvements made to our proactive detection technology on comments made on Facebook.</p>
	Violence and Incitement	0.02%	0.03%	0.03%	0.03%	





**Question 1:** How safe is the platform for consumers?

**Question 2:** How safe is the platform for advertisers?

**Authorized Metric:** Prevalence

Prevalence shows the potential of violating content actually being seen. Prevalence considers all the views of content on Facebook and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2022	Q3 2022	Q2 2022	Q1 2022	
<b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b>	Adult Nudity and Sexual Activity	<b>0.06%</b>	<b>0.05%</b>	0.04%	0.04%	<p>Adult Nudity and Sexual Activity: Prevalence increased in Q4 2022 due to bugs within our systems that have now been mitigated in addition to adjustments made to our proactive detection technology.</p> <p>Violence and Incitement: Prevalence decreased in Q4 2022 due to improvements made to our proactive detection technology on comments made on Facebook.</p> <p>Bullying and Harassment: Prevalence decreased due to AI improvements.</p> <p>Child Endangerment: Nudity and Physical Abuse and Child Endangerment: Sexual Exploitation. We cannot estimate prevalence for these right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.</p>
	Violence & Incitement	<b>0.02%</b>	<b>0.03%</b>	0.03%	0.03%	
	Violent and Graphic Content	<b>0.04%</b>	<b>0.04%</b>	0.04%	0.03-0.04%	
	Bullying and Harassment	<b>0.07-0.08%</b>	<b>0.08%</b>	0.08-0.09%	0.09-0.10%	
	Child Nudity and Sexual Exploitation	N/A	N/A	N/A	N/A	
	Child Endangerment: Nudity and Physical Abuse	N/A	N/A	N/A	N/A	
	Child Endangerment: Sexual Exploitation	N/A	N/A	N/A	N/A	
	Suicide and Self-Injury	<b>Less than 0.05%</b>	<b>Less than 0.05%</b>	Less than 0.05%	Less than 0.05%	
	Restricted Goods and Services: Firearms	<b>Less than 0.05%</b>	<b>Less than 0.05%</b>	Less than 0.05%	Less than 0.05%	



**Question 1:** How safe is the platform for consumers?

**Question 2:** How safe is the platform for advertisers?

**Authorized Metric:** Prevalence

Prevalence shows the potential of violating content actually being seen. Prevalence considers all the views of content on Facebook and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2022	Q3 2022	Q2 2022	Q1 2022	
Death, Injury or Military Conflict	Violent and Graphic Content	0.04%	0.04%	0.04%	0.03-0.04%	Violence and Incitement: Prevalence decreased in Q4 2022 due to improvements made to our proactive detection technology on comments made on Facebook.
	Violence and Incitement	0.02%	0.03%	0.03%	0.03%	
	Suicide and Self Injury	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	We do not report prevalence of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	
Hate speech & acts of aggression	Hate Speech	0.02%	0.02%	0.02%	0.02%	Dangerous Organizations: Organized Hate: We cannot estimate prevalence for Organized Hate right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.  Bullying and Harassment: Prevalence decreased due to AI improvements.
	Dangerous Organizations: Terrorism	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
	Dangerous Organizations: Organized Hate	N/A	N/A	N/A	N/A	
	Bullying and Harassment	0.07-0.08%	0.08%	0.08-0.09%	0.09-0.10%	



**Question 1:** How safe is the platform for consumers?

**Question 2:** How safe is the platform for advertisers?

**Authorized Metric:** Prevalence

Prevalence shows the potential of violating content actually being seen. Prevalence considers all the views of content on Facebook and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2022	Q3 2022	Q2 2022	Q1 2022	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	0.02%	0.02%	0.02%	0.02%	Bullying and Harassment: Prevalence decreased due to AI improvements.
	Bullying and Harassment	0.07-0.08%	0.08%	0.08-0.09%	0.09-0.10%	
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Restricted Goods and Services: Drugs	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
Spam or Harmful Content	Spam	N/A	N/A	N/A	N/A	We cannot estimate this metric right now. We are working on new methods to measure the prevalence of spam on Facebook. Our existing methods for measuring prevalence, which rely on people to manually review samples of content, do not fully capture this type of highly adversarial violation, which includes deceptive behavior as well as content. Spammy behavior, such as excessive resharing, cannot always be detected by reviewing the content alone. We are working on ways to review and classify spammers' behavior to build a comprehensive picture.
Terrorism	Dangerous Organizations: Terrorism	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	Dangerous Organizations: Organized Hate: We cannot estimate prevalence for Organized Hate right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
	Dangerous Organizations: Organized Hate	N/A	N/A	N/A	N/A	
Debated Sensitive Social Issue	Hate Speech	0.02%	0.02%	0.02%	0.02%	Bullying and Harassment: Prevalence decreased due to AI improvements.
	Bullying and Harassment	0.07-0.08%	0.08%	0.08-0.09%	0.09-0.10%	





**Question 3: How Effective is the Platform in Enforcing Safety Policy?**  
**Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate**

Violating content acted upon and removed by Facebook

GARM Category	Relevant Policy	Latest Period				Previous Period			
		Q4 2022		Q3 2022		Q2 2022		Q1 2022	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
<b>Adult &amp; Explicit Sexual Content</b>	Adult Nudity and Sexual Activity	29,200,000	94.1%	29,400,000	96.9%	38,400,000	97.2%	31,000,000	96.7%
	Child Endangerment: Nudity and Physical Abuse	2,500,000	98.7%	2,300,000	97.5%	1,900,000	97.3%	2,100,000	97.8%
	Child Endangerment: Sexual Exploitation	25,100,000	99.2%	30,100,000	99.5%	20,400,000	99.1%	16,500,000	96.4%
<b>Arms &amp; Ammunition</b>	Restricted Goods and Services: Firearms	1,600,000	87.1%	1,400,000	94.8%	1,600,000	94.4%	1,200,000	94.6%
	Violence & Incitement	13,100,000	87.3%	14,400,000	94.3%	19,300,000	98.2%	21,700,000	98.1%
<b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b>	Adult Nudity and Sexual Activity	29,200,000	94.1%	29,400,000	96.9%	38,400,000	97.2%	31,000,000	96.7%
	Violence & Incitement	13,100,000	87.3%	14,400,000	94.3%	19,300,000	98.2%	21,700,000	98.1%
	Violent and Graphic Content	15,500,000	98.1%	23,200,000	99.1%	45,900,000	99.5%	26,100,000	99.5%
	Bullying and Harassment	6,400,000	61.0%	6,600,000	67.8%	8,200,000	76.7%	9,500,000	67.0%
	Child Nudity and Sexual Exploitation	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Child Endangerment: Nudity and Physical Abuse	2,500,000	98.7%	2,300,000	97.5%	1,900,000	97.3%	2,100,000	97.8%
	Child Endangerment: Sexual Exploitation	25,100,000	99.2%	30,100,000	99.5%	20,400,000	99.1%	16,500,000	96.4%
	Suicide and Self-Injury	3,100,000	97.3%	5,600,000	98.6%	11,300,000	99.1%	6,800,000	98.8%
	Restricted Goods and Services: Firearms	1,600,000	87.1%	1,400,000	94.8%	1,600,000	94.4%	1,200,000	94.6%
<b>Death, Injury or Military Conflict</b>	Violent and Graphic Content	15,500,000	98.1%	23,200,000	99.1%	45,900,000	99.5%	26,100,000	99.5%
	Violence and Incitement	13,100,000	87.3%	14,400,000	94.3%	19,300,000	98.2%	21,700,000	98.1%
	Suicide and Self Injury	3,100,000	97.3%	5,600,000	98.6%	11,300,000	99.1%	6,800,000	98.8%
<b>Online piracy</b>	Intellectual Property: Copyright	N/A*	N/A*	N/A*	N/A*	2,800,000	87.9%	2,400,000	87.5%
	Intellectual Property: Counterfeit	N/A*	N/A*	N/A*	N/A*	1,200,000	95.3%	1,200,000	98.6%
	Intellectual Property: Trademark	N/A*	N/A*	N/A*	N/A*	360,300	N/A	413.3k	N/A







**Question 3: How Effective is the Platform in Enforcing Safety Policy?**  
**Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate**

Violating content acted upon and removed by Facebook

GARM Category	Relevant Policy	Latest Period				Previous Period			
		Q4 2022		Q3 2022		Q2 2022		Q1 2022	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
<b>Hate speech &amp; acts of aggression</b>	Hate Speech	11,000,000	81.9%	10,600,000	90.2%	13,500,000	95.6%	15,100,000	95.6%
	Dangerous Organizations: Terrorism	9,900,000	98.5%	16,700,000	99.1%	13,500,000	98.9%	16,100,000	98.8%
	Dangerous Organizations: Organized Hate	1,100,000	93.3%	1,200,000	94.3%	2,300,000	96.9%	2,500,000	96.9%
	Bullying and Harassment	6,400,000	61.0%	6,600,000	67.8%	8,200,000	76.7%	9,500,000	67.0%
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b>	Hate Speech	11,000,000	81.9%	10,600,000	90.2%	13,500,000	95.6%	15,100,000	95.6%
	Bullying and Harassment	6,400,000	61.0%	6,600,000	67.8%	8,200,000	76.7%	9,500,000	67.0%
<b>Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol</b>	Restricted Goods and Services: Drugs	5,400,000	99.1%	4,100,000	98.3%	3,900,000	98.1%	3,300,000	97.7%
<b>Spam or Harmful Content</b>	Spam	1,800,000,000	96.7%	1,400,000,000	98.5%	734,200,000	99.2%	1,800,000,000	99.7%
<b>Terrorism</b>	Dangerous Organizations: Terrorism	9,900,000	98.5%	16,700,000	99.1%	13,500,000	98.9%	16,100,000	98.8%
	Dangerous Organizations: Organized Hate	1,100,000	93.3%	1,200,000	94.3%	2,300,000	96.9%	2,500,000	96.9%
<b>Debated Sensitive Social Issue</b>	Hate Speech	11,000,000	81.9%	10,600,000	90.2%	13,500,000	95.6%	15,100,000	95.6%
	Bullying and Harassment	6,400,000	61.0%	6,600,000	67.8%	8,200,000	76.8%	9,500,000	67.0%



**Question 3:** How Effective is the Platform in Enforcing Safety Policy?  
**Authorized Metric:** Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Facebook

**Note Across Policies:** We consistently refine our methodology for this report in an effort to improve the metrics we provide. In Q4 2022, we updated how we calculate proactive rate. As a result of these improvements, we're seeing several shifts in the proactive rates across different areas. This methodology update only changes how we measure the proactive rate metric, but not our approach to proactively identifying violating content.

GARM Category	Relevant Policy	Commentary
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	
	Child Endangerment: Nudity and Physical Abuse	
	Child Endangerment: Sexual Exploitation	
Arms & Ammunition	Restricted Goods and Services: Firearms	
	Violence & Incitement	<p>Our actions against content that incites violence decreased from 19.3 million in Q2 2022 to 14.4 million in Q3 2022 after our improved AI technology was able to better recognize language and emojis used in jest between friends. As we improved our accuracy on this front, our proactive rate for actioning this content decreased from 98.2% to 94.3% in Q3 2022.</p> <p>In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.</p>



**Question 3: How Effective is the Platform in Enforcing Safety Policy?**  
**Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate**

Violating content acted upon and removed by Facebook

GARM Category	Relevant Policy	Commentary
<b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b>	Adult Nudity and Sexual Activity	
	Violence & Incitement	Our actions against content that incites violence decreased from 19.3 million in Q2 2022 to 14.4 million in Q3 2022 after our improved AI technology was able to better recognize language and emojis used in jest between friends. As we improved our accuracy on this front, our proactive rate for actioning this content decreased from 98.2% to 94.3% in Q3 2022.  In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
	Violent and Graphic Content	
	Bullying and Harassment	For bullying and harassment-related content, our proactive rate decreased in Q3 2022 from 76.7% in Q2 2022 to 67.8% in Q3 2022 on Facebook. This decrease was due to improved accuracy in our technologies (and a bug in our system that is now resolved).  In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
	Child Nudity and Sexual Exploitation	
	Child Endangerment: Nudity and Physical Abuse	
	Child Endangerment: Sexual Exploitation	
	Suicide and Self-Injury	
	Restricted Goods and Services: Firearms	
	<b>Death, Injury or Military Conflict</b>	Violent and Graphic Content
Violence and Incitement		Our actions against content that incites violence decreased from 19.3 million in Q2 2022 to 14.4 million in Q3 2022 after our improved AI technology was able to better recognize language and emojis used in jest between friends. As we improved our accuracy on this front, our proactive rate for actioning this content decreased from 98.2% to 94.3% in Q3 2022.  In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
Suicide and Self Injury		



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

**Authorized Metric:** Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Facebook

GARM Category	Relevant Policy	Commentary
Online piracy	Intellectual Property: Copyright	<p>We report these metrics in biannual report broken down monthly. These numbers reflect the total amount of content that was removed based on an IP report. On Facebook, this includes everything from individual posts, photos, videos or advertisements to profiles, Pages, groups and events.</p> <p>We do not report Proactive Rate for Intellectual Property: Trademark because we do not remove this type of violating content proactively. These violations require notice from the trademark owner.</p>
	Intellectual Property: Counterfeit	
	Intellectual Property: Trademark	
Hate speech & acts of aggression	Hate Speech	<p>Our actions against hate speech-related content decreased from 13.5 million to 10.6 million in Q3 2022 on Facebook because we improved the accuracy of our AI technology. We've done this by leveraging data from past user appeals to identify posts that could have been removed by mistake without appropriate cultural context. For example, now we can better recognize humorous terms of endearment used between friends, or better detect words that may be considered offensive or inappropriate in one context but not another. As we improved this accuracy, our proactive detection rate for hate speech also decreased from 95.6% to 90.2% in Q3 2022.</p> <p>In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.</p>
	Dangerous Organizations: Terrorism	On Facebook in Q3 2022, we took action on 16.7 million pieces of content related to terrorism, an increase from 13.5 million in Q2. This increase was because non-violating videos were added incorrectly to our media-matching technology banks and were removed (though they were eventually restored).
	Dangerous Organizations: Organized Hate	
	Bullying and Harassment	<p>For bullying and harassment-related content, our proactive rate decreased in Q3 2022 from 76.7% in Q2 2022 to 67.8% in Q3 2022 on Facebook. This decrease was due to improved accuracy in our technologies (and a bug in our system that is now resolved).</p> <p>In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.</p>
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	<p>Our actions against hate speech-related content decreased from 13.5 million to 10.6 million in Q3 2022 on Facebook because we improved the accuracy of our AI technology. We've done this by leveraging data from past user appeals to identify posts that could have been removed by mistake without appropriate cultural context. For example, now we can better recognize humorous terms of endearment used between friends, or better detect words that may be considered offensive or inappropriate in one context but not another. As we improved this accuracy, our proactive detection rate for hate speech also decreased from 95.6% to 90.2% in Q3 2022.</p> <p>In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.</p>
	Bullying and Harassment	<p>For bullying and harassment-related content, our proactive rate decreased in Q3 2022 from 76.7% in Q2 2022 to 67.8% in Q3 2022 on Facebook. This decrease was due to improved accuracy in our technologies (and a bug in our system that is now resolved).</p> <p>In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.</p>
Illegal Drugs /Tobacco/e-cigarettes / Vaping /Alcohol	Restricted Goods and Services: Drugs	On Facebook in Q3 2022, we took action on 4.1 million pieces of drug content, an increase from 3.9 million in Q2 2022, due to improvements made to our proactive detection technology.



**Question 3: How Effective is the Platform in Enforcing Safety Policy?**  
**Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate**

Violating content acted upon and removed by Facebook

GARM Category	Relevant Policy	Commentary
<b>Spam or Harmful Content</b>	Spam	On Facebook in Q3, we took action on 1.4 billion pieces of spam content, an increase from 734 million in Q2, due to an increased number of adversarial spam incidents in August.
<b>Terrorism</b>	Dangerous Organizations: Terrorism	On Facebook in Q3 2022, we took action on 16.7 million pieces of content related to terrorism, an increase from 13.5 million in Q2. This increase was because non-violating videos were added incorrectly to our media-matching technology banks and were removed (though they were eventually restored).
	Dangerous Organizations: Organized Hate	
<b>Debated Sensitive Social Issue</b>	Hate Speech	Our actions against hate speech-related content decreased from 13.5 million to 10.6 million in Q3 2022 on Facebook because we improved the accuracy of our AI technology. We've done this by leveraging data from past user appeals to identify posts that could have been removed by mistake without appropriate cultural context. For example, now we can better recognize humorous terms of endearment used between friends, or better detect words that may be considered offensive or inappropriate in one context but not another. As we improved this accuracy, our proactive detection rate for hate speech also decreased from 95.6% to 90.2% in Q3 2022.  In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
	Bullying and Harassment	For bullying and harassment-related content, our proactive rate decreased in Q3 2022 from 76.7% in Q2 2022 to 67.8% in Q3 2022 on Facebook. This decrease was due to improved accuracy in our technologies (and a bug in our system that is now resolved).  In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.



## Question 4: How does the platform perform at correcting mistakes?

### Authorized Metric: Appeals, Reinstatements

Content that is acted upon and then appealed by users, and the decision to reinstate it

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2022			Q3 2022			Q2 2022			Q1 2022			
		Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	2,100,000	614,300	143,900	2,300,000	576,900	123,400	2,500,000	462,900	71,200	274,300	47,000	239,600	
	Child Endangerment: Nudity and Physical Abuse	94,700	13,600	541,700	85,000	14,600	29,900	61,700	11,300	18,700	4,000	700	21,200	
	Child Endangerment: Sexual Exploitation	23,000	2,600	75,800	414,200	4,000	205,300	404,000	1,400	15,900	800	100	687,800	
Arms & Ammunition	Restricted Goods and Services: Firearms	132,400	20,400	17,200	153,100	27,500	9,700	149,900	31,800	31,700	74,200	12,800	67,500	
	Violence and Incitement	2,900,000	370,400	19,300	3,500,000	454,100	5,200	4,500,000	554,500	6,900	756,000	77,700	402,800	



## Question 4: How does the platform perform at correcting mistakes?

### Authorized Metric: Appeals, Reinstatements

Content that is acted upon and then appealed by users, and the decision to reinstate it

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2022			Q3 2022			Q2 2022			Q1 2022			
		Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	2,100,000	614,300	143,900	2,300,000	576,900	123,400	2,500,000	462,900	71,200	274,300	47,000	239,600	
	Violence and Incitement	2,900,000	370,400	19,300	3,500,000	454,100	5,200	4,500,000	554,500	6,900	756,000	77,700	402,800	
	Violent and Graphic Content	29,300	6,100	23,100	54,600	7,100	2,400	53,800	8,400	34,700	4,500	800	12,000	
	Bullying and Harassment	1,300,000	239,500	34,500	1,500,000	288,600	14,400	1,900,000	514,600	28,800	736,000	114,700	333,400	
	Child Endangerment: Nudity and Physical Abuse	94,700	13,600	541,700	85,000	14,600	29,900	61,700	11,300	18,700	4,000	700	21,200	
	Child Endangerment: Sexual Exploitation	23,000	2,600	75,800	414,200	4,000	205,300	404,000	1,400	15,900	800	100	687,800	
	Suicide and Self-Injury	146,800	54,000	116,600	258,600	105,900	171,900	461,000	186,600	595,100	6,100	1,700	343,900	
	Restricted Goods and Services: Firearms	132,400	20,400	17,200	153,100	27,500	9,700	149,900	31,800	31,700	74,200	12,800	67,500	



## Question 4: How does the platform perform at correcting mistakes?

### Authorized Metric: Appeals, Reinstatements

Content that is acted upon and then appealed by users, and the decision to reinstate it

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2022		Q3 2022		Q2 2022		Q1 2022						
		Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	
Death, Injury or Military Conflict	Violent and Graphic Content	29,300	6,100	23,100	54,600	7,100	2,400	53,800	8,400	34,700	4,500	800	12,000	
	Violence and Incitement	2,900,000	370,400	19,300	3,500,000	454,100	5,200	4,500,000	554,500	6,900	756,000	77,700	402,800	
	Suicide and Self Injury	146,800	54,000	116,600	258,600	105,900	171,900	461,000	186,600	595,100	6,100	1,700	343,900	
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	We do not report content appealed and reinstated of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
Hate speech & acts of aggression	Hate Speech	2,200,000	198,200	12,800	2,400,000	184,400	7,100	2,700,000	237,900	11,800	586,700	48,700	218,200	
	Dangerous Organizations: Terrorism	373,700	83,300	207,200	332,000	60,300	4,000,000	531,000	61,900	24,400	33,800	5,400	408,300	
	Dangerous Organizations: Organized Hate	178,800	38,500	11,000	157,200	43,500	17,600	204,900	60,200	9,100	34,000	12,000	219,600	
	Bullying and Harassment	1,300,000	239,500	34,500	1,500,000	288,600	14,400	1,900,000	514,600	28,800	736,000	114,700	333,400	





## Question 4: How does the platform perform at correcting mistakes?

### Authorized Metric: Appeals, Reinstatements

Content that is acted upon and then appealed by users, and the decision to reinstate it

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2022			Q3 2022			Q2 2022			Q1 2022			
		Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	2,200,000	198,200	12,800	2,400,000	184,400	7,100	2,700,000	237,900	11,800	586,700	48,700	218,200	
	Bullying and Harassment	1,300,000	239,500	34,500	1,500,000	288,600	14,400	1,900,000	514,600	28,800	736,000	114,700	333,400	
Illegal Drugs/Tobacco/e-cigarettes/Vaping /Alcohol	Restricted Goods and Services: Drugs	428,000	80,400	32,300	297,000	39,200	18,800	241,500	44,100	51,900	104,100	37,600	111,400	
Spam or Harmful Content	Spam	1,400,000	242,400	50,400	1,400,000	240,200	94,200,000	611,900	84,000	117,300,000	39,700	1,700	33,000,000	
Terrorism	Dangerous Organizations: Terrorism	373,700	83,300	207,200	332,000	60,300	4,000,000	531,000	61,900	24,400	33,800	5,400	408,300	
	Dangerous Organizations: Organized Hate	178,800	38,500	11,000	157,200	43,500	17,600	204,900	60,200	9,100	34,000	12,000	219,600	
Debated Sensitive Social Issue	Hate Speech	2,200,000	198,200	12,800	2,400,000	184,400	7,100	2,700,000	237,900	11,800	586,700	48,700	218,200	
	Bullying and Harassment	1,300,000	239,500	34,500	1,500,000	288,600	14,400	1,900,000	514,600	28,800	736,000	114,700	333,400	



**Question 1:** How safe is the platform for consumers?

**Question 2:** How safe is the platform for advertisers?

**Authorized Metric:** Prevalence

Prevalence shows the potential of violating content actually being seen. Prevalence considers all the views of content on Instagram and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2022	Q3 2022	Q2 2022	Q1 2022	
<b>Adult &amp; Explicit Sexual Content</b>	Adult Nudity and Sexual Activity	0.03-0.04%	0.02-0.03%	0.02-0.03%	0.02-0.03%	Child Endangerment: Nudity and Physical Abuse and Child Endangerment: Sexual Exploitation. We cannot estimate prevalence for these right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
	Child Endangerment: Sexual Exploitation	N/A	N/A	N/A	N/A	
	Child Endangerment: Nudity and Physical Abuse	N/A	N/A	N/A	N/A	
<b>Arms &amp; Ammunition</b>	Restricted Goods and Services: Firearms	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
	Violence and Incitement	0.02-0.03%	0.02%	0.01-0.02%	0.01-0.02%	



**Question 1:** How safe is the platform for consumers?

**Question 2:** How safe is the platform for advertisers?

**Authorized Metric:** Prevalence

Prevalence shows the potential of violating content actually being seen. Prevalence considers all the views of content on Instagram and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2022	Q3 2022	Q2 2022	Q1 2022	
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	0.03-0.04%	0.02-0.03%	0.02-0.03%	0.02-0.03%	Violent and Graphic Content: Prevalence increased in Q4 2022 due to bugs within our systems that have now been mitigated in addition to adjustments made to our proactive detection technology
	Violence and Incitement	0.02-0.03%	0.02%	0.01-0.02%	0.01-0.02%	
	Violent and Graphic Content	0.03%	0.02%	0.01-0.02%	0.01-0.02%	
	Bullying and Harassment	0.05-0.06%	0.04-0.05%	0.04-0.05%	0.05-0.06%	
	Child Endangerment: Nudity and Physical Abuse	N/A	N/A	N/A	N/A	
	Child Endangerment: Sexual Exploitation	N/A	N/A	N/A	N/A	
	Suicide and Self-Injury	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
	Restricted Goods and Services: Firearms	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	



**Question 1:** How safe is the platform for consumers?

**Question 2:** How safe is the platform for advertisers?

**Authorized Metric:** Prevalence

Prevalence shows the potential of violating content actually being seen. Prevalence considers all the views of content on Instagram and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2022	Q3 2022	Q2 2022	Q1 2022	
Death, Injury or Military Conflict	Violent and Graphic Content	0.03%	0.02%	0.01%-0.02%	0.01%-0.02%	Violent and Graphic Content: Prevalence increased in Q4 2022 due to bugs within our systems that have now been mitigated in addition to adjustments made to our proactive detection technology
	Violence and Incitement	0.02-0.03%	0.02%	0.01-0.02%	0.01-0.02%	
	Suicide and Self Injury	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	We do not report prevalence of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	
Hate speech & acts of aggression	Hate Speech	0.01-0.02%	0.01-0.02%	0.01-0.02%	0.02%	Dangerous Organizations: Organized Hate-We cannot estimate prevalence for organized hate right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
	Dangerous Organizations: Terrorism	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
	Dangerous Organizations: Organized Hate	N/A	N/A	N/A	N/A	
	Bullying and Harassment	0.05-0.06%	0.04-0.05%	0.04-0.05%	0.05-0.06%	



**Question 1:** How safe is the platform for consumers?

**Question 2:** How safe is the platform for advertisers?

**Authorized Metric:** Prevalence

Prevalence shows the potential of violating content actually being seen. Prevalence considers all the views of content on Instagram and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2022	Q3 2022	Q2 2022	Q1 2022	
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b>	Hate Speech	0.01-0.02%	0.01-0.02%	0.01-0.02%	0.02%	
	Bullying and Harassment	0.05-0.06%	0.04-0.05%	0.04-0.05%	0.05-0.06%	
<b>Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol</b>	Restricted Goods and Services: Drugs	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
<b>Spam or Harmful Content</b>	Spam	N/A	N/A	N/A	N/A	We cannot estimate this metric right now. We are working on new methods to measure the prevalence of spam on Instagram. Our existing methods for measuring prevalence, which rely on people to manually review samples of content, do not fully capture this type of highly adversarial violation, which includes deceptive behavior as well as content. Spammy behavior, such as excessive resharing, cannot always be detected by reviewing the content alone. We are working on ways to review and classify spammers' behavior to build a comprehensive picture.
<b>Terrorism</b>	Dangerous Organizations: Terrorism	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	Dangerous Organizations: Organized Hate -- We cannot estimate prevalence for organized hate right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
	Dangerous Organizations: Organized Hate	N/A	N/A	N/A	N/A	
<b>Debated Sensitive Social Issue</b>	Hate Speech	0.01-0.02%	0.01-0.02%	0.01-0.02%	0.02%	
	Bullying and Harassment	0.05-0.06%	0.04-0.05%	0.04-0.05%	0.05-0.06%	



**Question 3: How Effective is the Platform in Enforcing Safety Policy?**  
**Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate**

Violating content acted upon and removed by Instagram

GARM Category	Relevant Policy	Latest Period				Previous Period			
		Q4 2022		Q3 2022		Q2 2022		Q1 2022	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	10,800,000	95.5%	11,400,000	95.7%	10,300,000	94.3%	10,400,000	94.0%
	Child Endangerment: Sexual Exploitation	9,700,000	99.6%	1,300,000	96.6%	1,200,000	94.9%	1,500,000	92.5%
	Child Endangerment: Nudity and Physical Abuse	620,700	97.6%	1,000,000	96.8%	480,500	93.4%	600,000	93.8%
Arms & Ammunition	Restricted Goods and Services: Firearms	155,300	95.0%	238,200	94.0%	214,800	93.6%	151,000	92.2%
	Violence and Incitement	5,300,000	97.6%	4,500,000	97.5%	3,700,000	97.0%	2,700,000	95.4%
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	10,800,000	95.5%	11,400,000	95.7%	10,300,000	94.3%	10,400,000	94.0%
	Violence and Incitement	5,300,000	97.6%	4,500,000	97.5%	3,700,000	97.0%	2,700,000	95.4%
	Violent and Graphic Content	6,100,000	98.8%	6,900,000	99.1%	10,100,000	99.3%	6,100,000	99.0%
	Bullying and Harassment	5,000,000	85.4%	6,100,000	84.3%	6,100,000	87.4%	7,000,000	83.8%
	Child Endangerment: Nudity and Physical Abuse	620,700	97.6%	1,000,000	96.8%	480,500	93.4%	600,700	93.8%
	Child Endangerment: Sexual Exploitation	9,700,000	99.6%	1,300,000	96.6%	1,200,000	94.9%	1,500,000	92.5%
	Suicide and Self-Injury	5,000,000	98.6%	5,700,000	98.4%	6,400,000	98.4%	5,100,000	98.0%
	Restricted Goods and Services: Firearms	155,300	95.0%	238,200	94.0%	214,800	93.6%	151,000	92.2%
Death, Injury or Military Conflict	Violent and Graphic Content	6,100,000	98.8%	6,900,000	99.1%	10,100,000	99.3%	6,100,000	99.0%
	Violence and Incitement	5,300,000	97.6%	4,500,000	97.5%	3,700,000	97.0%	2,700,000	95.4%
	Suicide and Self Injury	5,000,000	98.6%	5,700,000	98.4%	6,400,000	98.4%	5,100,000	98.0%
Online piracy	Intellectual Property: Copyright	N/A*	N/A*	N/A*	N/A*	1,100,000	88.0%	885,800	88.4%
	Intellectual Property: Counterfeit	N/A*	N/A*	N/A*	N/A*	321,100	84.7%	317,300	80.5%
	Intellectual Property: Trademark	N/A*	N/A*	N/A*	N/A*	147,400	N/A	139,100	N/A

\* We release the H2 2022 figures in May, and they are not yet available at the time of this report publishing



**Question 3: How Effective is the Platform in Enforcing Safety Policy?**  
**Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate**

Violating content acted upon and removed by Instagram

GARM Category	Relevant Policy	Latest Period				Previous Period			
		Q4 2022		Q3 2022		Q2 2022		Q1 2022	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
<b>Hate speech &amp; acts of aggression</b>	Hate Speech	4,700,000	93.9%	4,300,000	93.7%	3,800,000	91.2%	3,400,000	89.6%
	Dangerous Organizations: Terrorism	1,200,000	93.9%	2,200,000	96.7%	1,900,000	93.3%	1,500,000	86.3%
	Dangerous Organizations: Organized Hate	376,200	89.7%	373,000	86.3%	450,000	87.7%	481,300	88.9%
	Bullying and Harassment	5,000,000	85.4%	6,100,000	84.3%	6,100,000	87.4%	7,000,000	83.8%
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b>	Hate Speech	4,700,000	93.9%	4,300,000	93.7%	3,800,000	91.2%	3,400,000	89.6%
	Bullying and Harassment	5,000,000	85.4%	6,100,000	84.3%	6,100,000	87.4%	7,000,000	83.8%
<b>Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol</b>	Restricted Goods and Services: Drugs	3,100,000	98.8%	2,500,000	97.1%	1,900,000	96.8%	1,800,000	96.0%
<b>Spam or Harmful Content</b>	Spam	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
<b>Terrorism</b>	Dangerous Organizations: Terrorism	1,200,000	93.9%	2,200,000	96.7%	1,900,000	93.3%	1,500,000	86.3%
	Dangerous Organizations: Organized Hate	376,200	89.7%	373,000	86.3%	450,000	87.7%	481,300	88.9%
<b>Debated Sensitive Social Issue</b>	Hate Speech	4,700,000	93.9%	4,300,000	93.7%	3,800,000	91.2%	3,400,000	89.6%
	Bullying and Harassment	5,000,000	85.4%	6,100,000	84.3%	6,100,000	87.4%	7,000,000	83.8%



**Question 3:** How Effective is the Platform in Enforcing Safety Policy?  
**Authorized Metric:** Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Instagram

**Note Across Policies:** We consistently refine our methodology for this report in an effort to improve the metrics we provide. In Q4 2022, we [updated](#) how we calculate proactive rate. As a result of these improvements, we're seeing several shifts in the proactive rates across different areas. This methodology update only changes how we measure the proactive rate metric, but not our approach to proactively identifying violating content.

GARM Category	Relevant Policy	Commentary
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	
	Child Endangerment: Sexual Exploitation	
	Child Endangerment: Nudity and Physical Abuse	
Arms & Ammunition	Restricted Goods and Services: Firearms	
	Violence and Incitement	In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.





**Question 3: How Effective is the Platform in Enforcing Safety Policy?**  
**Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate**

Violating content acted upon and removed by Instagram

GARM Category	Relevant Policy	Commentary
<b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b>	Adult Nudity and Sexual Activity	
	Violence and Incitement	In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
	Violent and Graphic Content	
	Bullying and Harassment	In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
	Child Endangerment: Nudity and Physical Abuse	
	Child Endangerment: Sexual Exploitation	
	Suicide and Self-Injury	
	Restricted Goods and Services: Firearms	
<b>Death, Injury or Military Conflict</b>	Violent and Graphic Content	
	Violence and Incitement	In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
	Suicide and Self Injury	
<b>Online piracy</b>	Intellectual Property: Copyright	We report these metrics in biannual report broken down monthly These numbers reflect the total amount of content that was removed based on an IP report.  We do not report Proactive Rate for Intellectual Property: Trademark because we do not remove this type of violating content proactively. These violations require notice from the trademark owner.
	Intellectual Property: Counterfeit	
	Intellectual Property: Trademark	



**Question 3: How Effective is the Platform in Enforcing Safety Policy?**  
**Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate**

Violating content acted upon and removed by Instagram

GARM Category	Relevant Policy	Commentary
<b>Hate speech &amp; acts of aggression</b>	Hate Speech	In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
	Dangerous Organizations: Terrorism	On Instagram in Q3 2022, we took action on 2.2 million pieces of content related to terrorism, from 1.9 million in Q2, due to non-violating videos added incorrectly to our media-matching technology banks and were removed (though they were eventually restored).
	Dangerous Organizations: Organized Hate	
	Bullying and Harassment	For bullying and harassment-related content, our proactive rate decreased in Q3 2022 from 87.4% to 84.3% on Instagram. This decrease was due to improved accuracy in our technologies (and a bug in our system that is now resolved).  In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b>	Hate Speech	In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
	Bullying and Harassment	For bullying and harassment-related content, our proactive rate decreased in Q3 2022 from 87.4% to 84.3% on Instagram. This decrease was due to improved accuracy in our technologies (and a bug in our system that is now resolved).  In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
<b>Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol</b>	Restricted Goods and Services: Drugs	On Instagram in Q3 2022, we took action on 2.5 million pieces of drug content, an increase from 1.9 million, due to improvements in our proactive detection technology.
<b>Spam or Harmful Content</b>	Spam	
<b>Terrorism</b>	Dangerous Organizations: Terrorism	On Instagram in Q3 2022, we took action on 2.2 million pieces of content related to terrorism, from 1.9 million in Q2, due to non-violating videos added incorrectly to our media-matching technology banks and were removed (though they were eventually restored).
	Dangerous Organizations: Organized Hate	
<b>Debated Sensitive Social Issue</b>	Hate Speech	In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.
	Bullying and Harassment	For bullying and harassment-related content, our proactive rate decreased in Q3 2022 from 87.4% to 84.3% on Instagram. This decrease was due to improved accuracy in our technologies (and a bug in our system that is now resolved).  In Q4 2022, we updated our cross-problem AI system, combining several models so that we're consolidating learnings across hate speech, bullying and harassment and violence and incitement. This and other continued improvements or adjustments to proactive detection technology, in many instances, lead to improved accuracy.



## Question 4: How does the platform perform at correcting mistakes?

### Authorized Metric: Appeals, Reinstatements

Content that is acted upon and then appealed by users, and the decision to reinstate it

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2022			Q3 2022			Q2 2022			Q1 2022			
		Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	393,400	101,900	126,900	590,400	140,900	75,100	867,200	185,800	80,400	0	0	183,800	
	Child Endangerment: Sexual Exploitation	5,800	100	2,400	3,500	200	7,100	4,100	200	400	0	20	154,200	
	Child Endangerment: Nudity and Physical Abuse	16,000	2,000	4,900	36,000	4,100	6,400	29,200	3,800	5,900	0	0	10,700	
Arms & Ammunition	Restricted Goods and Services: Firearms	5,000	800	1,000	19,000	9,100	5,600	21,600	13,800	6,200	0	0	15,400	
	Violence and Incitement	252,900	38,400	30,000	318,900	45,500	10,200	327,000	54,300	10,200	0	0	53,100	



## Question 4: How does the platform perform at correcting mistakes?

### Authorized Metric: Appeals, Reinstatements

Content that is acted upon and then appealed by users, and the decision to reinstate it

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2022			Q3 2022			Q2 2022			Q1 2022			
		Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	393,400	101,900	126,900	590,400	140,900	75,100	867,200	185,800	80,400	0	0	183,800	
	Violence and Incitement	252,900	38,400	30,000	318,900	45,500	10,200	327,000	54,300	10,200	0	0	53,100	
	Violent and Graphic Content	9,600	4,200	4,800	25,300	5,900	2,200	157,900	19,700	1,100,000	0	0	31,400	
	Bullying and Harassment	465,700	121,700	32,800	680,500	96,500	11,100	853,500	168,600	18,200	0	0	215,800	
	Child Endangerment: Nudity and Physical Abuse	16,000	2,000	4,900	36,000	4,100	6,400	29,200	3,800	5,900	0	0	10,700	
	Child Endangerment: Sexual Exploitation	5,800	100	2,400	3,500	200	7,100	4,100	200	400	0	20	154,200	
	Suicide and Self-Injury	86,400	24,100	28,200	102,300	46,300	24,400	177,600	79,100	40,000	0	0	49,400	
	Restricted Goods and Services: Firearms	5,000	800	1,000	19,000	9,100	5,600	21,600	13,800	6,200	0	0	15,400	



## Question 4: How does the platform perform at correcting mistakes?

### Authorized Metric: Appeals, Reinstatements

Content that is acted upon and then appealed by users, and the decision to reinstate it

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2022			Q3 2022			Q2 2022			Q1 2022			
		Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	
Death, Injury or Military Conflict	Violent and Graphic Content	9,600	4,200	4,800	25,300	5,900	2,200	157,900	19,700	1,100,000	0	0	31,400	
	Violence and Incitement	252,900	38,400	30,000	318,900	45,500	10,200	327,000	54,300	10,200	0	0	53,100	
	Suicide and Self Injury	86,400	24,100	28,200	102,300	46,300	24,400	177,600	79,100	40,000	0	0	49,400	
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	We do not report content appealed and reinstated of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
Hate speech & acts of aggression	Hate Speech	229,000	21,200	18,300	352,800	36,400	9,800	397,400	47,800	14,200	0	0	56,700	
	Dangerous Organizations: Terrorism	34,600	4,200	4,000	56,700	12,700	915,500	99,400	18,000	4,600	0	0	84,900	
	Dangerous Organizations: Organized Hate	23,900	5,300	4,600	26,100	6,800	9,300	46,900	14,700	6,200	0	0	31,200	
	Bullying and Harassment	465,700	121,700	32,800	680,500	96,500	11,100	853,500	168,600	18,200	0	0	215,800	



## Question 4: How does the platform perform at correcting mistakes?

### Authorized Metric: Appeals, Reinstatements

Content that is acted upon and then appealed by users, and the decision to reinstate it

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2022			Q3 2022			Q2 2022			Q1 2022			
		Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	Appealed	Reinstated with appeal	Reinstated without appeal	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	229,000	21,200	18,300	352,800	36,400	9,800	397,400	47,800	14,200	0	0	56,700	
	Bullying and Harassment	465,700	121,700	32,800	680,500	96,500	11,100	853,500	168,600	18,200	0	0	215,800	
Illegal Drugs/Tobacco/e-cigarettes/Vaping /Alcohol	Restricted Goods and Services: Drugs	105,700	12,100	13,100	120,300	10,700	4,100	160,000	22,400	5,200	0	0	45,000	
Spam or Harmful Content	Spam	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
Terrorism	Dangerous Organizations: Terrorism	34,600	4,200	4,000	56,700	12,700	915,500	99,400	18,000	4,600	0	0	84,900	
	Dangerous Organizations: Organized Hate	23,900	5,300	4,600	26,100	6,800	9,300	46,900	14,700	6,200	0	0	31,200	
Debated Sensitive Social Issue	Hate Speech	229,000	21,200	18,300	352,800	36,400	9,800	397,400	47,800	14,200	0	0	56,700	
	Bullying and Harassment	465,700	121,700	32,800	680,500	96,500	11,100	853,500	168,600	18,200	0	0	215,800	



Transparency is at the heart of Twitter’s commitment to serve the public conversation. It underpins the development of our policies, our products, and our partnerships – all of which drive towards making Twitter a safer place for people and brands.

Our commitment to transparency continues post acquisition at Twitter 2.0, and we’re evolving the platform to be more transparent than ever before. This includes enhanced transparency in content moderation, open-sourcing the Twitter algorithm, providing first- and third-party insights on the health of the platform, enhanced brand safety controls and reporting for our advertisers, and other important initiatives.

Our [21st Transparency Report](#) shares data from January 1st to June 30th, 2022. While this update is in a different format from our past updates to the [Twitter Transparency Center](#), it continues to share familiar insights that Twitter has reported for years, including account and content removals, by policy vertical. Due to data retention policies, we weren’t able to report on the full suite of metrics that we’ve published and included in past Aggregated Measurement Reports, but we strongly believed that it was important to share as much information as possible for H1 2022. We remain committed to transparency and will continue to identify opportunities to share regular, useful reporting with our users and our customers.

Over the reporting period, Twitter required users to remove 6,586,109 pieces of content that violated the Twitter Rules, an increase of 29% from H2 2021. We took enforcement action on 5,096,272 accounts during this period (a 20% increase), and 1,618,855 accounts were suspended for violating the Twitter Rules (a 28% increase).

Around the world, Twitter received approximately 53,000 legal requests to remove content from governments during the reporting period. Twitter’s compliance rate for these requests varied by requester country. The top requesting countries were Japan, South Korea, Turkey and India.

Twitter received over 16,000 government information requests for user data from over 85 countries during the reporting period. Disclosure rates vary by requester country. The top five requesting countries seeking account information in H1 2022 were India, the United States, France, Japan, and Germany.

We continue to improve our detection and enforcement capabilities over time, aggressively removing bad actors and illegal content and preventing the amplification and distribution of other toxic content.

We intend to share more about our path forward for transparency reporting later this year. In the meantime, we will continue to share insights into Twitter’s work towards serving the public conversation from [@TwitterSafety](#).



## Question 1: How safe is the platform for consumers?

### Next Best Measure: Content Removals

GARM Category	Relevant Twitter Policy	Latest Period – H1 2022	Previous Period – H2 2021	Commentary
		Content Removals	Impressions	
Adult & explicit sexual content	Non-consensual nudity	Over the reporting period, Twitter required users to remove 6,586,109 pieces of content that violated the Twitter Rules, an increase of 29% from H2 2021.	In total, impressions on these violative Tweets accounted for less than 0.1% of all impressions for all Tweets during that time period.	Due to data retention limitations, Twitter is not able to report on violative impressions for H1 2022 in its totality. However, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets during the period in H1 2022 for which data was available.
	Sensitive media			
	Child sexual exploitation			
Arms & ammunition	Illegal or certain regulated goods or services			
Crime & harmful acts to individuals and society, human right violations	Violence			
	Abuse/harassment			
Death, injury or military conflict	Promoting suicide or self-harm			
Online piracy	Copyright			
	Trademark			
Hate speech & acts of aggression	Hateful conduct			
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media			
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services			
Spam or harmful content	Private information			
	Impersonation			
	Platform manipulation			
Terrorism	Terrorism/violent extremism			
Debated sensitive social issues	N/A			
Other	Civic integrity			
	COVID-19 misleading information			





## Question 2: How safe is the platform for advertisers?

**Next best measure:** Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2022			Previous Period – H2 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
<b>Adult &amp; explicit sexual content</b>	Non-consensual nudity	68,714	16,670	1,524,067	28,836	8,141	60,816	
	Sensitive media	1,315,670	150,757	1,352,155	1,143,064	118,356	1,149,829	
	Child sexual exploitation	696,015	691,704	11,927	599,523	596,997	6,796	
<b>Arms &amp; ammunition</b>	Illegal or certain regulated goods or services	399,297	249,328	1,365,341	224,185	119,508	571,902	
<b>Crime &amp; harmful acts to individuals and society, human right violations</b>	Violence	28,753	19,838	35,240	61,358	41,386	70,229	
	Abuse/harassment	1,083,788	96,284	1,524,067	940,679	82,971	1,344,061	
<b>Death, injury or military conflict</b>	Promoting suicide or self-harm	439,555	11,776	547,377	408,143	10,197	509,776	
<b>Online piracy</b>	Copyright	Notices Issued: Not reported	Accounts Affected: Not reported	Tweets Withheld: Not reported	Notices Issued: 146,906	Accounts Affected: 623,576	Tweets Withheld: 161,983	
	Trademark	Trademark Notices: Not reported			Trademark Notices: 26,274			
<b>Hate speech &amp; acts of aggression</b>	Hateful conduct	1,085,651	111,056	1,527,442	902,169	104,565	1,293,178	
<b>Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b>	Sensitive media	1,315,670	150,757	1,352,155	1,143,064	118,356	1,149,829	
<b>Illegal drugs/tobacco/e-cigarettes/vaping/alcohol</b>	Illegal or certain regulated goods or services	399,297	249,328	1,365,341	224,185	119,508	571,902	
<b>Spam or harmful content</b>	Private information	45,844	2,536	78,357	34,181	2,563	62,537	
	Impersonation	266,034	249,572	19,798	181,644	169,396	15,275	
	Platform manipulation	Anti-Spam Challenges Issued: Not reported			Anti-Spam Challenges Issued: 133,266,534			
<b>Terrorism</b>	Terrorism/violent extremism	30,616	30,616	0	33,694	33,693	1	
<b>Debated sensitive social issues</b>	N/A							
<b>Other</b>	Civic integrity	Not reported			93	4	102	
	COVID-19 misleading information	Not reported			24,012	1,376	30,190	



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

#### Authorized Metric: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2022			Previous Period – H2 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Adult & explicit sexual content	Non-consensual nudity	68,714	16,670	1,524,067	28,836	8,141	60,816	There was a 138% increase in the number of accounts actioned for violations of our non-consensual nudity policy during this reporting period.
	Sensitive media	1,315,670	150,757	1,352,155	1,143,064	118,356	1,149,829	There was a 15% increase in the number of accounts actioned for violations of our sensitive media policy during this reporting period.  We removed a total of 1,352,155 unique pieces of content under our Sensitive Media policy during this period, an 18% increase since our last report.
	Child sexual exploitation	696,015	691,704	11,927	599,523	596,997	6,796	There was a 16% increase in the number of accounts actioned for violations of our child sexual exploitation policy during this reporting period.  We do not tolerate child sexual exploitation on Twitter. When we are made aware of child sexual exploitation media, including links to images of or content promoting child exploitation, the material will be removed from the site without further notice and reported to The National Center for Missing & Exploited Children ("NCMEC"). People can report content that appears to violate the <a href="#">Twitter Rules regarding Child Sexual Exploitation</a> via our <a href="#">web form</a> .
Arms & ammunition	Illegal or certain regulated goods or services	399,297	249,328	1,365,341	224,185	119,508	571,902	There was a 78% increase in the number of accounts actioned for violations of our illegal or certain regulated goods or services policy during this reporting period.
Crime & harmful acts to individuals and society, human right violations	Violence	28,753	19,838	35,240	61,358	41,386	70,229	There was a 53% decrease in the number of accounts actioned for violations of our violence policies during this reporting period.  Our policies prohibit sharing content that threatens violence against an individual or a group of people. We also prohibit the glorification of violence. 19,838 accounts were suspended and we took action on 35,240 unique pieces of content during this reporting period.
	Abuse/harassment	1,083,788	96,284	1,524,067	940,679	82,971	1,344,061	There was a 15% increase in the number of accounts actioned for violations of our abuse policy during this reporting period.  Under our Abusive Behavior policy, we prohibit content that harasses or intimidates, or is otherwise intended to shame or degrade others. We took action on 1,083,788 accounts during this reporting period.



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

#### Authorized Metric: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2022			Previous Period – H2 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Death, injury or military conflict	Promoting suicide or self-harm	439,555	11,776	547,377	408,143	10,197	509,776	There was an 8% increase in the number of accounts actioned for violations of our suicide or self-harm policy during this reporting period.
Online piracy	Copyright	Notices Issued: <b>Not reported</b>	Accounts Affected: <b>Not reported</b>	Tweets Withheld: <b>Not reported</b>	Notices Issued: <b>146,906</b>	Accounts Affected: <b>623,576</b>	Tweets Withheld: <b>161,983</b>	Although Twitter is not disclosing data around trademark and copyright removals this reporting period, Twitter continues to receive and respond to trademark and copyright notices.
	Trademark	Trademark Notices: <b>Not reported</b>			Trademark Notices: <b>26,274</b>			We carefully review each report received under our <a href="#">trademark policy</a> , and follow up with the reporter as appropriate, such as in cases of apparent fair use. We may take action on reported content if it is using another's trademark in a manner that may mislead others about its business affiliation.
Hate speech & acts of aggression	Hateful conduct	1,085,651	111,056	1,527,442	902,169	104,565	1,293,178	There was a 20% increase in the number of accounts actioned for violations of our hateful conduct policy during this reporting period.
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	1,315,670	150,757	1,352,155	1,143,064	118,356	1,149,829	There was a 15% increase in the number of accounts actioned for violations of our sensitive media policy during this reporting period.  We removed a total of 1,352,155 unique pieces of content under our Sensitive Media policy during this period, an 18% decrease since our last report.



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

#### Authorized Metric: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2022			Previous Period – H2 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Spam or harmful content	Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	399,297	249,328	1,365,341	224,185	119,508	571,902	<p>There was a 78% increase in the number of accounts actioned for violations of our illegal or certain regulated goods or services policy during this reporting period.</p> <p>Due to continued refinement of enforcement guidelines, we saw a 109% increase in accounts suspended under this policy, representing a total of 249,328 accounts.</p>
	Private information	45,844	2,536	78,357	34,181	2,563	62,537	<p>There was a 34% increase in the number of accounts actioned for violations of our private information policy during this reporting period.</p> <p>45,844 accounts and 78,357 unique pieces of content were actioned under this policy.</p>
	Impersonation	266,034	249,572	19,798	181,644	169,396	15,275	<p>There was a 46% increase in the number of accounts actioned for violations of our impersonation policy during this time period.</p> <p>This reporting period, we suspended 249,572 accounts, a 47% increase.</p>
	Platform manipulation	Anti-Spam Challenges Issued: <b>Not reported</b>			Anti-Spam Challenges Issued: <b>133,266,534</b>			<p>One way we fight manipulation and spam at scale is to use anti-spam challenges to confirm whether an authentic account holder is in control of accounts engaged in suspicious activity. For example, we may require the account holder to verify a phone number or email address, or to complete a CAPTCHA test. These challenges are simple for authentic account owners to solve, but difficult (or costly) for spammers to complete. Accounts which fail to complete a challenge within a specified period of time may be suspended.</p> <p>We share this metric at our discretion, and may not share this metric every reporting period.</p>



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

#### Authorized Metric: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2022			Previous Period – H2 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Terrorism	Terrorism/violent extremism	30,616	30,616	0	33,694	33,693	1	<p>There was a 9% decrease in the number of accounts actioned for violations of our terrorism/violent extremism policy during this reporting period.</p> <p>We suspended 30,616 unique accounts for violations of the policy during this reporting period. Our current methods of surfacing potentially violating content for review, in addition to our proactive detection, include leveraging the shared industry hash database supported by the Global Internet Forum to Counter Terrorism (GIFCT).</p>
Debated sensitive social issues	N/A							
Other	Civic integrity	Not reported			93	4	102	
	COVID-19 misleading information	Not reported			24,012	1,376	30,190	



### Question 4: How does the platform perform at correcting mistakes?

Not submitted

GARM Category	Relevant Twitter Policy	Commentary
Adult & explicit sexual content	Non-consensual nudity	Twitter does not report appeals data at this time.
	Sensitive media	
	Child sexual exploitation	
Arms & ammunition	Illegal or certain regulated goods or services	
Crime & harmful acts to individuals and society, human right violations	Violence	
	Abuse/harassment	
Death, injury or military conflict	Promoting suicide or self-harm	
Online piracy	Copyright	
	Trademark	
Hate speech & acts of aggression	Hateful conduct	
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services	
Spam or harmful content	Private information	
	Impersonation	
	Platform manipulation	
Terrorism	Terrorism/violent extremism	
Debated sensitive social issues	N/A	
Other	Civic integrity	
	COVID-19 misleading information	





### About TikTok's Community Guidelines Enforcement Reports

TikTok is a global entertainment platform fueled by the creativity of our diverse community. We strive to foster a fun and inclusive environment where people can create, find community, and be entertained. To maintain that environment, we take action upon content and accounts that violate our Community Guidelines or Terms of Service and regularly publish information about these actions to hold ourselves accountable to our community.

TikTok uses a combination of innovative technology and people to identify, review, and action content that violates our policies. This report provides quarterly insights into the volume and nature of content and accounts removed from our platform.

### Evolving Our Approach

In the past, we've scaled moderation by casting a wider net for review and working to catch as much violative content as possible. While this generally increased the number of videos we were removing, it doesn't measure our overarching safety goals of prioritizing egregious content, minimizing overall views, and ensuring accurate and consistent decisions. As our community has continued to grow and express themselves—including through new experiences like longer videos, LIVE, and TikTok Now—our approach to content moderation has evolved as well. We're increasingly focused on preventing overall harm across features while building a fair, accurate experience for our creators.

As a result, in recent months we've started refining our approach to better prioritize accuracy, minimize views of violative content, and remove egregious content quickly. We've upgraded the systems that route content for review, so that they better incorporate a video's severity of harm (based on the type of potential violation) and expected reach (based on an account's following) when determining whether to remove it, escalate for human review, or take a different course of action.

We're leveraging measures like age-restricted features, ineligibility for recommendation, and our new Content Levels system more frequently and transparently, to help ensure content reaches appropriate audiences. And our proactive technology is driving down the amount of content that needs human review, as it grows more sophisticated at catching things like spam accounts at sign-up, or duplicative content.

### Our Q4 Transparency Report

The impact of these changes is already being reflected in the data from our Q4 Community Guidelines Enforcement report. For example, total content removals dipped as we made more low-harm content ineligible for the For You feed rather than fully removing it. At the same time, the proportion of that content which was accurately removed by automation increased as our systems became more precise. This fluctuation is within our expected range—we consistently remove 1% percent or less of published content for being violative—and is part of a concerted effort to make our growing community safer and to foster a more consistent experience for our creators.

It's possible such metric fluctuations will continue as we continue to evolve our systems over the coming year. For example, we've recently made significant additional improvements such as introducing a new account enforcement system and comprehensively refreshing our Community Guidelines with new policy groupings that future Enforcement Reports will follow. We're also continuing to refine moderation processes behind the scenes, such as by specializing more content moderation teams around areas of expertise.

### Looking Forward

There's no finish line when it comes to keeping people safe, and our latest report and continued safety improvements reflect our unwavering commitment to the safety and well-being of our community. We look forward to sharing more about our ongoing work to safeguard our platform.

In the meantime, you can read up on other transparency efforts at our refreshed Transparency Center: [www.tiktok.com/transparency](http://www.tiktok.com/transparency).



# How our policies map to the GARM Brand Safety Floor

GARM Category	Relevant Policy
Hate speech & acts of aggression	[Policy] <b>Hateful behavior</b> [Sub-policy] 1. Hateful ideology; 2. Attacks and slurs on the basis of protected attributes
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	[Policy] <b>Hateful behavior</b> [Sub-policy] 1. Attacks and slurs on the basis of protected attributes
	[Policy] <b>Violent and Graphic Content</b>
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	[Policy] <b>Illegal activities and regulated goods</b> [Sub-policy] 1. Drugs, controlled substances, alcohol, and tobacco
Spam or Harmful Content	[Policy] <b>Integrity and authenticity</b> [Sub-policy] 1. Spam and fake engagement
	[Policy] <b>Illegal activities and regulated goods</b> [Sub-policy] 1. Frauds and scams
Terrorism	[Policy] <b>Violent extremism</b> [Sub-policy] 1. Threats and incitement to violence; 2. Violent extremist organizations and individuals
	[Policy] <b>Illegal activities and regulated goods</b> [Sub-policy] 1. Criminal Activities; 2. Weapons
Debated Sensitive Social Issue	[Policy] <b>Hateful behavior</b> [Sub-policy] 1. Hateful ideology; 2. Attacks and slurs on the basis of protected attributes
	[Policy] <b>Suicide, self-harm, and disordered eating</b> [Sub-policy] 1. Disordered eating; 2. Suicide and self-harm;
Adult & Explicit Sexual Content	[Policy] <b>Minor Safety</b> [Sub-policy] 1. Sexual exploitation of minors; 2. Nudity and sexual activity involving minors
	[Policy] <b>Adult nudity and sexual activities</b> [Sub-policy] 1. Sexual exploitation; 2. Nudity and sexual activity involving adults
Arms & Ammunition	[Policy] <b>Illegal activities and regulated goods</b> [Sub-policy] 1. Weapons
Crime & Harmful acts to individuals and Society, Human Right Violations	[Policy] <b>Illegal activities and regulated goods</b> [Sub-policy] 1. Criminal Activities; 2. Frauds and Scams; 3. Privacy, personal data, and personally identifiable information
	[Policy] <b>Harassment and Bullying</b> [Sub-policy] 1. Abusive Behavior; 2. Sexual Harassment; 3. Threats of hacking, doxxing, and blackmail
	[Policy] <b>Hateful Behavior</b> [Sub-policy] 1. Attacks and slurs on the basis of protected attributes
	[Policy] <b>Dangerous Acts and Challenges</b> [Sub-policy] 1. Dangerous Acts and Challenges*
	[Policy] <b>Suicide, self-harm, and disordered eating</b> [Sub-policy] 1. Suicide and self-harm
Death, Injury or Military Conflict	[Policy] <b>Minor Safety</b> [Sub-policy] 1. Grooming behavior; 2. Physical and psychological harm of minors
	[Policy] <b>Violent and graphic content</b>
Online piracy	[Policy] <b>Suicide, self-harm, and disordered eating</b> [Sub-policy] 1.. Suicide and self-harm;
	[Policy] <b>Integrity and authenticity</b> [Sub-policy] 1. Intellectual property violations
Misinformation	[Policy] <b>Integrity and authenticity</b> [Sub-policy] 1. Harmful Misinformation

**Note:** The above policies reflect our previous Community Guidelines, which we reported against in Q3-Q4 2022. Our guidelines have since been expanded and we will update our mapping to include our new policy areas moving forward.

\*Dangerous Acts and Challenges became separate TikTok Issue Policy effective Q4 2022.





**Question 1:** How safe is the platform for consumers?

**Next best measure:** Overall videos removed, and overall removal rates

Removals	Latest Period		Previous Period	
	Q4 2022	Q3 2022	Q2 2022	Q1 2022
Total Videos Removed	85,680,819	110,954,663	113,809,300	102,305,516
Videos Removed by Automation	46,836,047	53,287,839	48,011,571	34,726,592
Percentage of videos removed proactively before being reported by a user:	96.2%	96.5%	95.9%	95.1%
Percentage of videos removed before receiving any views:	84.7%	89.5%	90.5%	90.0%
Percentage of videos removed within 24 hours of being posted:	91.2%	92.7%	93.7%	93.7%





# Question 1: How safe is the platform for consumers?

**Next best measure:** Percentage of videos removed by policy violation

Volume of videos removed by policy violation, as a percentage of total videos removed

TikTok Policy	Latest Period		Previous Period		Applicable GARM Categories
	Q4 2022	Q3 2022	Q2 2022	Q1 2022	
Adult nudity and sexual activities	12.8%	10.7%	10.7%	11.3%	Adult & Explicit Sexual Content
Harassment and bullying	6.1%	5.9%	5.7%	6.0%	Crime & Harmful acts to individuals and Society, Human Right Violations
Hateful behavior	2.3%	2.0%	1.7%	1.6%	Hate speech & acts of aggression; Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust; Debated Sensitive Social Issue; Crime & Harmful acts to individuals and Society, Human Right Violations
Illegal activities and regulated goods	27.4%	21.0%	21.2%	21.8%	Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol; Spam or Harmful Content; Terrorism; Arms & Ammunition; Crime & Harmful acts to individuals and Society, Human Right Violations
Integrity and authenticity	0.9%	0.7%	0.7%	0.6%	Spam or Harmful Content; Online piracy; Misinformation
Minor safety	33.3%	42.9%	43.7%	41.7%	Adult & Explicit Sexual Content; Crime & Harmful acts to individuals and Society, Human Right Violations
Dangerous Acts and Challenges*	5.0%	6.5%	6.1%	6.7%	Crime & Harmful acts to individuals and Society
Suicide, self-harm, and disordered eating	2.8%				Debated Sensitive Social Issue, Human Right Violations; Death, Injury or Military Conflict
Violent and graphic content	8.0%	9.3%	9.3%	9.6%	Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust; Death, Injury or Military Conflict
Violent extremism	1.4%	1.0%	0.9%	0.7%	Terrorism

\*Dangerous Acts and Challenges became separate TikTok Issue Policy effective Q4 2022.



## Question 2: How safe is the platform for advertisers?

Not submitted

Relevant Policy	Latest Period	Previous Period	Commentary
Adult nudity and sexual activities			<p>We do not currently report on the prevalence of ad adjacency to violative content. Because our ads are 100% share of voice (full-screen), we offer a 0% on-screen adjacency environment. Additionally, all videos adjacent to (before and after) advertisements are reviewed by technology and human moderators and must be eligible for recommendation.</p> <p>Additionally, we provide a proprietary first-party solution, the TikTok Inventory Filter, which allows advertisers to have more control over the type of videos that are adjacent to their ads. Available in 30+ markets, the TikTok Inventory Filter offers 3 tiers of user-generated video inventory - Full, Standard and Limited - for advertisers to choose from to run before and after their ads. Powered by advanced machine learning technology, the TikTok Inventory Filter's tiers are populated based on analysis of four levels of risk across 17 content categories - all of which are informed by TikTok Community Guidelines, Terms of Service and Intellectual Property Guidelines as well as the GARM Brand Safety Floor and Brand Suitability Framework.</p>
Harassment and bullying			
Hateful behavior			
Illegal activities and regulated goods			
Integrity and authenticity			
Minor safety			
Dangerous Acts and Challenges			
Suicide, self-harm, and disordered eating			
Violent and graphic content			
Violent extremism			

\*Dangerous Acts and Challenges became separate TikTok Issue Policy effective Q4 2022.



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

#### Next Best Measure: Removal rates by policy

Percentage of violating videos removed proactively, before any views and within 24 hours, by policy.

TikTok Policy	Latest Period						Previous Period						Applicable GARM Categories
	Q4 2022			Q3 2022			Q2 2022			Q1 2022			
	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	
Adult nudity and sexual activities	92.6%	78.8%	90.6%	93.3%	82.1%	90.9%	90.2%	80.3%	90.1%	89.4%	78.3%	88.9%	Adult & Explicit Sexual Content
Harassment and bullying	84.5%	65.3%	81.4%	83.8%	70.7%	83.10%	82.4%	71.4%	83.5%	78.9%	69.4%	83.9%	Crime & Harmful acts to individuals and Society, Human Right Violations
Hateful behavior	87.1%	73.2%	83.7%	85.5%	74.8%	83.8%	81.1%	72.4%	83.5%	77.0%	68.3%	82.2%	Hate speech & acts of aggression; Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust; Debated Sensitive Social Issue; Crime & Harmful acts to individuals and Society, Human Right Violations
Illegal activities and regulated goods	97.6%	86.0%	92.9%	97.7%	91.2%	93.9%	97.6%	93.1%	94.9%	97.1%	93.8%	95.6%	Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol; Spam or Harmful Content; Terrorism; Arms & Ammunition; Crime & Harmful acts to individuals and Society, Human Right Violations
Integrity and authenticity	91.5%	69.5%	71.3%	92.0%	77.2%	83.5%	89.1%	74.7%	83.9%	83.6%	60.8%	71.9%	Spam or Harmful Content; Online piracy; Misinformation
Minor safety	98.5%	91.1%	93.0%	98.5%	93.9%	94.4%	98.4%	95.4%	95.7%	98.1%	95.5%	95.9%	Adult & Explicit Sexual Content; Crime & Harmful acts to individuals and Society, Human Right Violations
Dangerous Acts and Challenges*	96.0%	70.5%	82.2%	96.5%	84.2%	88.4%	96.2%	85.4%	88.5%	95.3%	86.1%	90.4%	Crime & Harmful acts to individuals and Society
Suicide, self-harm, and disordered eating	98.1%	93.7%	95.6%										Debated Sensitive Social Issue;, Human Right Violations; Death, Injury or Military Conflict
Violent and graphic content	96.9%	85.3%	91.8%	97.7%	91.2%	93.6%	97.3%	91.5%	94.0%	96.4%	89.9%	94.3%	Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust; Death, Injury or Military Conflict
Violent extremism	94.0%	79.6%	88.3%	94.1%	85.5%	89.0%	92.8%	82.2%	85.6%	91.4%	83.9%	88.4%	Terrorism

**Note:** Proactive removal means identifying and removing a video before it's reported. Removal within 24 hours means removing the video within 24 hours of it being posted on our platform.

\*Dangerous Acts and Challenges became separate TikTok Issue Policy effective Q4 2022.



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

#### Next Best Measure: Removal rates by sub-policy

Percentage of violating videos removed proactively, before any views and within 24 hours, by sub-policy.

TikTok Issue Policy	TikTok Sub- Policy	Latest Period						Previous Period					
		Q4 2022			Q3 2022			Q2 2022			Q1 2022		
		Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours
Adult nudity and sexual activities	Nudity and sexual activity involving adults	86.4%	56.8%	79.8%	89.4%	73.5%	85.0%	87.4%	74.4%	85.8%	87.7%	74.1%	85.4%
	Sexual exploitation	85.8%	67.8%	90.2%	90.6%	72.5%	89.9%	81.0%	64.7%	85.7%	80.4%	64.7%	87.2%
Harassment and bullying	Abusive behavior	82.4%	60.8%	79.0%	82.1%	68.2%	81.5%	81.2%	70.3%	82.7%	78.2%	68.9%	83.6%
	Sexual harassment	74.7%	49.3%	74.0%	76.2%	55.8%	75.6%	73.3%	53.7%	74.7%	68.9%	52.2%	75.8%
	Threats of hacking, doxxing, and blackmail	87.4%	64.5%	77.6%	87.2%	72.0%	79.8%	87.8%	79.4%	81.3%	87.5%	81.4%	86.6%
Hateful behavior	Attacks and slurs on the basis of protected attributes	90.2%	78.6%	86.2%	89.4%	78.9%	85.2%	85.0%	75.9%	84.5%	81.7%	72.0%	83.5%
	Hateful ideology	77.8%	58.4%	76.6%	76.1%	62.4%	77.3%	72.7%	63.8%	79.1%	68.9%	60.6%	78.4%

**Note:** Only videos that have been reviewed by moderators are included in the sub-policy dashboard. Our minor safety policies aim to promote the highest standard of safety and well-being for teens. The “sexual activity involving minors” sub-policy prohibits a broad range of content, including “minors in minimal clothing” and “sexually explicit dancing”; these two categories represent the majority of content removed under that sub-policy. Child Sexual Abuse Material (CSAM) is reported separately.



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

#### Next Best Measure: Removal rates by sub-policy

Percentage of violating videos removed proactively, before any views and within 24 hours, by policy.

TikTok Issue Policy	TikTok Sub- Policy	Latest Period						Previous Period					
		Q4 2022			Q3 2022			Q2 2022			Q1 2022		
		Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours
Illegal activities and regulated goods	Criminal activities	77.0%	30.9%	57.1%	82.90%	49.40%	69.40%	87.4%	66.1%	80.2%	82.8%	62.4%	80.3%
	Drugs, controlled substances, alcohol, and tobacco	93.0%	63.0%	77.0%	93.90%	77.40%	83.00%	94.4%	83.5%	86.9%	94.2%	86.6%	89.5%
	Frauds and scams	87.2%	52.7%	76.7%	88.50%	63.70%	80.50%	83.1%	62.8%	81.0%	79.2%	68.9%	83.9%
	Gambling	89.5%	44.5%	75.0%	93.20%	70.60%	83.50%	95.2%	77.5%	85.9%	94.3%	82.5%	88.8%
	Privacy, personal data, and personally identifiable information	97.6%	69.1%	84.5%	98.50%	91.10%	91.00%	98.8%	94.3%	93.5%	98.4%	95.4%	95.2%
	Weapons	96.1%	74.5%	85.3%	96.80%	88.00%	87.30%	97.4%	92.6%	90.2%	97.2%	93.9%	93.0%
Integrity and authenticity	Harmful Misinformation	89.1%	69.8%	59.9%	78.80%	55.30%	59.60%	70.0%	39.1%	57.5%	66.1%	37.6%	50.6%
	Spam and fake engagement	89.9%	55.0%	74.0%	87.30%	60.60%	76.40%	82.9%	59.9%	76.6%	84.4%	66.1%	81.3%



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

**Next Best Measure:** Removal rates by sub-policy.

Percentage of violating videos removed proactively, before any views and within 24 hours, by policy.

TikTok Issue Policy	TikTok Sub- Policy	Latest Period						Previous Period					
		Q4 2022			Q3 2022			Q2 2022			Q1 2022		
		Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours
Minor safety	Grooming behavior	96.2%	73.9%	76.9%	96.6%	83.4%	84.9%	96.7%	88.4%	90.0%	96.6%	90.6%	92.7%
	Harmful activities by minors	96.9%	81.1%	83.8%	96.5%	86.6%	85.5%	97.1%	90.7%	88.0%	96.8%	92.9%	91.3%
	Nudity and sexual activity involving minors	96.7%	79.7%	83.0%	97.0%	87.2%	86.5%	96.9%	91.4%	89.9%	96.8%	92.6%	91.8%
	Physical and psychological harm of a minor	95.4%	65.0%	76.5%	96.4%	83.9%	84.9%	96.7%	87.8%	86.5%	96.2%	88.2%	90.1%
	Sexual exploitation of minors	93.1%	82.1%	88.7%	95.1%	88.3%	92.5%	93.2%	85.8%	90.7%	90.6%	82.5%	90.3%
Dangerous Acts and Challenges*	Dangerous acts and challenges	93.5%	51.7%	70.8%	94.5%	72.3%	79.8%	95.0%	78.8%	84.5%	94.4%	82.6%	88.8%
Suicide, self-harm, and disordered eating	Disordered eating	85.4%	64.5%	75.2%	86.6%	71.0%	79.3%	86.7%	75.2%	82.5%	82.0%	70.9%	82.2%
	Suicide and self-harm	96.7%	88.6%	91.0%	96.7%	91.6%	91.1%	97.1%	93.4%	91.5%	97.1%	94.0%	94.0%

\*Dangerous Acts and Challenges became separate TikTok Issue Policy effective Q4 2022.



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

**Next Best Measure:** Removal rates by sub-policy.

Percentage of violating videos removed proactively, before any views and within 24 hours, by policy.

TikTok Issue Policy	TikTok Sub- Policy	Latest Period						Previous Period					
		Q4 2022			Q3 2022			Q2 2022			Q1 2022		
		Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours	Proactive Removal Rate	Removal Rate Before Any Views	Removal Rate Within 24 Hours
Violent and graphic content	Violent and graphic content	96.9%	85.3%	91.8%	88.9%	77.5%	86.0%	95.9%	86.9%	89.7%	95.0%	86.4%	91.8%
Violent extremism	Threats and incitement to violence	87.6%	55.2%	81.0%	93.3%	82.6%	85.1%	84.7%	75.6%	84.3%	80.9%	69.5%	82.4%
	Violent extremist organizations and individuals	93.0%	78.5%	86.3%	97.7%	91.2%	93.6%	93.0%	85.9%	88.8%	93.3%	87.0%	89.8%





### Question 3: How Effective is the Platform in Enforcing Safety Policy?

**Authorized Metric:** Removal of Violating Accounts

	Total accounts removed	Accounts suspected to be under the age of 13 removed	Fake accounts removed	Other accounts removed	Total accounts removed	Accounts suspected to be under the age of 13 removed	Other accounts removed	Other accounts removed	Total accounts removed	Accounts suspected to be under the age of 13 removed	Fake accounts removed	Other accounts removed	Total accounts removed	Accounts suspected to be under the age of 13 removed	Fake accounts removed	Other accounts removed
Adult nudity and sexual activities	We removed a total of <b>76,178,122</b> accounts for violating Community Guidelines or Terms of Service.	<b>17,877,316</b>	<b>54,453,610</b>	<b>3,847,196</b>	We removed a total of <b>76,436,381</b> accounts for violating Community Guidelines or Terms of Service.	<b>19,690,699</b>	<b>50,963,108</b>	<b>5,782,574</b>	We removed a total of <b>59,430,082</b> accounts for violating Community Guidelines or Terms of Service.	<b>20,575,056</b>	<b>33,632,058</b>	<b>5,222,968</b>	We removed a total of <b>44,438,988</b> accounts for violating Community Guidelines or Terms of Service.	<b>20,219,476</b>	<b>20,890,519</b>	<b>3,328,993</b>
Harassment and bullying																
Hateful behavior																
Illegal activities and regulated goods																
Integrity and authenticity																
Minor safety																
Dangerous Acts and Challenges																
Suicide, self-harm, and disordered eating																
Violent and graphic content																
Violent extremism																



### Question 4: How does the platform perform at correcting mistakes?

#### Authorized Metric: Appeals & Reinstatements

Content removed by TikTok, appealed by users and then restored by TikTok.

Relevant Policy	Q4 2022	Q3 2022	Q2 2022	Q1 2022	Commentary
Adult nudity and sexual activities	We reinstated <b>5,477,549</b> videos after they were appealed	We reinstated <b>6,937,997</b> videos after they were appealed	We reinstated <b>5,896,218</b> videos after they were appealed	We reinstated <b>5,025,536</b> videos after they were appealed	Content reinstatement represents figure across all Community Guidelines
Harassment and bullying					
Hateful behavior					
Illegal activities and regulated goods					
Integrity and authenticity					
Minor safety					
Dangerous Acts and Challenges					
Suicide, self-harm, and disordered eating					
Violent and graphic content					
Violent extremism					





Our mission is our guiding light in drafting our content policies: to bring everyone the inspiration to create a life they love. When it comes to advertising and brand safety on Pinterest, it's important to remember that Pinterest is personal media—not social media—so things are a little different around here. On Pinterest, there are more “public” discovery surfaces like the home feed, and more “personal” surfaces, like individual users’ boards and profiles. Importantly, ads only show up on discovery surfaces, including home feed, search, and related Pins.

Pinterest is for inspiration, and it's hard to feel inspired if you don't feel safe. That's why we've been deliberate about engineering a more positive place online—that includes what we don't permit on Pinterest. For example, we don't allow harmful misinformation, like the promotion of false cures for terminal illnesses. We also don't allow political campaign ads.

It's important to be clear: Pinterest is absolutely not a place for antagonistic, explicit, false or misleading, hateful, or violent content or behavior. We may block, limit the distribution of, or remove content and the accounts, individuals and groups that create or spread that content based on how much harm it poses.

\*\*\*

We work with outside experts and organizations to inform our policies and content moderation practices and continue to invest heavily in measures, like machine learning technology, to fight policy-violating content on our platform. Over the years we've made advancements in the ability to detect similar images in Pins, and this technology has been applied to our content moderation work to take action at scale in appropriate circumstances.

We started publishing a semi-annual transparency report in 2013, and in 2020 we expanded the report to include new information. Now, our semi-annual transparency report includes data on the actions we take to moderate user and merchant content on Pinterest beyond those requested by law enforcement and government agencies, such as the number of policy violations and deactivations. In our latest report, we've also introduced reporting on Violent Actors.

Our latest transparency report includes data from H2 2022, broken out by quarter (Q3: July–September 2022, Q4: October–December 2022). During this reporting period, we took steps to strengthen our commitment to not only tackling false and misleading content that could interfere with civic engagement but also sharing reliable information on where and how to vote. We also announced a first-of-its-kind global partnership with Headspace, a leading meditation app, to support hundreds of thousands of creators with a free 6-month subscription.

Our mission at Pinterest is to bring everyone the inspiration to create a life they love. Let's create a safer, more inspiring internet, together.



### Note on methodology

To understand how we approach content moderation, it's helpful to differentiate between two types of Pins: organic Pins and ads. Our [Community guidelines](#) apply to both.

Organic Pins include all Pins created and saved on Pinterest that are not promoted as ads. For example, this could include merchants' product Pins, which aren't always ads, and may appear organically to people who are searching for products on Pinterest. We have additional requirements, like that the Pin image and description must accurately represent the product, for [merchants](#) and their product Pins. All types of organic Pins are included in our transparency reports.

Ads are Pins that businesses pay to promote. We have additional policies for [advertisers](#) that hold ads and advertisers to even higher standards. Ad content policies are enforced differently than organic content and are not included in our transparency reports.

Reach is one of our key indicators of user experience. In this report, we've updated the labels we use to report this metric, but the underlying methodology has remained the same:

- **Seen by 0 people** means that no one saw that Pin in the reporting period
- **Seen by 1-9 people** means a Pin was seen by at least 1 and no more than 9 people in the reporting period
- **Seen by 10-100 people** means a Pin was seen by at least 10 and up to 100 people in the reporting period
- **Seen by >100 people** means a Pin was seen by more than 100 people in the reporting period

Much of the content on Pinterest has been saved repeatedly, meaning that the same image may appear in multiple Pins. So when it comes to reporting content moderation for organic Pins, we include the number of Pins deactivated as well as the number of distinct images deactivated to provide greater insight into our moderation practices.

Because we report boards and accounts deactivated separately—and to avoid double-counting deactivations—our count of distinct images and Pins deactivated does not include those from boards or user accounts that were deactivated.

The latest period of data encompasses Q3 and Q4 2022.



# Question 1: How safe is the platform for consumers?

## Next Best Measure: Reach<sup>1</sup> of Pins deactivated for violating policy

Pinterest does not report on prevalence and instead uses reach as a metric due to the nature of the platform.

Pinterest Policy	Latest Period								Previous Period							
	Q4 2022				Q3 2022				Q2 2022				Q1 2022			
	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people
Adult content	54%	36%	7%	3%	71%	24%	4%	2%	77%	19%	3%	1%	78%	18%	3%	1%
Adult sexual services	62%	28%	5%	5%	64%	29%	5%	2%	82%	14%	3%	0.9%	95%	5%	0.5%	0.1%
Child sexual exploitation (CSE) <sup>2</sup>	51%	35%	9%	4%	61%	29%	6%	3%	61%	30%	6%	3%	63%	28%	6%	3%
Civic misinformation	82%	16%	1%	1%	64%	30%	4%	2%	46%	47%	5%	3%	20%	67%	5%	8%
Climate Misinformation <sup>4</sup>	1%	72%	8%	19%	27%	60%	7%	6%	13%	81%	4%	2%	N/A	N/A	N/A	N/A
Conspiracy theories	71%	23%	2%	4%	76%	19%	2%	2%	84%	14%	1%	1%	57%	33%	5%	5%
Copyright <sup>3</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Dangerous goods and activities	89%	10%	0.6%	0.3%	85%	13%	0.9%	0.4%	64%	31%	3%	2%	72%	21%	4%	3%
Graphic violence and threats	58%	40%	0.5%	1%	48%	48%	1%	3%	67%	28%	2%	4%	65%	29%	4%	3%
Harassment and criticism	65%	34%	0.8%	0.7%	68%	27%	2%	2%	58%	35%	4%	2%	74%	21%	3%	3%
Hateful activities <sup>4</sup>	52%	35%	4%	9%	52%	37%	5%	6%	62%	26%	4%	8%	63%	22%	6%	9%

<sup>1</sup> Calculated by looking at each policy-violating Pin deactivated in a reporting period, then counting the number of unique users that saw each of those Pins during the reporting period for at least 1 second, before it was deactivated.

<sup>2</sup> CSE includes any content that might exploit or endanger minors. By sharing reach for CSE content, we are not implying in any way that harm to children is somehow lessened if fewer people see it. The content is violative and wrong, no matter how many people see it. We share the data only to be transparent in our efforts to remove CSE from our platform.

<sup>3</sup> We do not currently report on reach for Copyright.

<sup>4</sup> Reach for Hateful activities and Climate misinformation in Q4 2022 does not include Pins deactivated that we later determined to be false positives and subsequently reinstated. See commentary for more details.



# Question 1: How safe is the platform for consumers?

## Next Best Measure: Reach<sup>1</sup> of Pins deactivated for violating policy

Pinterest does not report on prevalence and instead uses reach as a metric due to the nature of the platform.

Pinterest Policy	Latest Period								Previous Period							
	Q4 2022				Q3 2022				Q2 2022				Q1 2022			
	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people
Medical misinformation	63%	28%	3%	7%	64%	31%	2%	3%	91%	8%	0.3%	0.4%	85%	13%	0.7%	1%
Self-injury and harmful behavior	74%	21%	4%	1%	77%	21%	2%	0.6%	93%	6%	0.6%	0.4%	71%	23%	3%	3%
Spam	33%	49%	18%	0.7%	76%	20%	3%	1%	89%	10%	0.7%	0.3%	60%	34%	4%	1%
Trademark <sup>3</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Violent Actors	63%	28%	6%	3%	73%	21%	4%	2%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

<sup>1</sup> Calculated by looking at each policy-violating Pin deactivated in a reporting period, then counting the number of unique users that saw each of those Pins during the reporting period for at least 1 second, before it was deactivated.

<sup>2</sup> CSE includes any content that might exploit or endanger minors. By sharing reach for CSE content, we are not implying in any way that harm to children is somehow lessened if fewer people see it. The content is violative and wrong, no matter how many people see it. We share the data only to be transparent in our efforts to remove CSE from our platform.

<sup>3</sup> We do not currently report on reach for Trademark.

<sup>4</sup> Reach for Hateful activities and Climate misinformation in Q4 2022 does not include Pins deactivated that we later determined to be false positives and subsequently reinstated. See commentary for more details.



## Question 2: How safe is the platform for advertisers?

### Next Best Measure: Reach<sup>1</sup> of Pins deactivated for violating policy

Pinterest does not report on prevalence and instead uses reach as a metric due to the nature of the platform.

Pinterest Policy	Latest Period								Previous Period							
	Q4 2022				Q3 2022				Q2 2022				Q1 2022			
	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people
Adult content	54%	36%	7%	3%	71%	24%	4%	2%	77%	19%	3%	1%	78%	18%	3%	1%
Adult sexual services	62%	28%	5%	5%	64%	29%	5%	2%	82%	14%	3%	0.9%	95%	5%	0.5%	0.1%
Child sexual exploitation (CSE) <sup>2</sup>	51%	35%	9%	4%	61%	29%	6%	3%	61%	30%	6%	3%	63%	28%	6%	3%
Civic misinformation	82%	16%	1%	1%	64%	30%	4%	2%	46%	47%	5%	3%	20%	67%	5%	8%
Climate Misinformation <sup>4</sup>	1%	72%	8%	19%	27%	60%	7%	6%	13%	81%	4%	2%	N/A	N/A	N/A	N/A
Conspiracy theories	71%	23%	2%	4%	76%	19%	2%	2%	84%	14%	1%	1%	57%	33%	5%	5%
Copyright <sup>3</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Dangerous goods and activities	89%	10%	0.6%	0.3%	85%	13%	0.9%	0.4%	64%	31%	3%	2%	72%	21%	4%	3%
Graphic violence and threats	58%	40%	0.5%	1%	48%	48%	1%	3%	67%	28%	2%	4%	65%	29%	4%	3%
Harassment and criticism	65%	34%	0.8%	0.7%	68%	27%	2%	2%	58%	35%	4%	2%	74%	21%	3%	3%
Hateful activities <sup>4</sup>	52%	35%	4%	9%	52%	37%	5%	6%	62%	26%	4%	8%	63%	22%	6%	9%

<sup>1</sup> Calculated by looking at each policy-violating Pin deactivated in a reporting period, then counting the number of unique users that saw each of those Pins during the reporting period for at least 1 second, before it was deactivated.

<sup>2</sup> CSE includes any content that might exploit or endanger minors. By sharing reach for CSE content, we are not implying in any way that harm to children is somehow lessened if fewer people see it. The content is violative and wrong, no matter how many people see it. We share the data only to be transparent in our efforts to remove CSE from our platform.

<sup>3</sup> We do not currently report on reach for Copyright.

<sup>4</sup> Reach for Hateful activities and Climate misinformation in Q4 2022 does not include Pins deactivated that we later determined to be false positives and subsequently reinstated. See commentary for more details.



## Question 2: How safe is the platform for advertisers?

### Next Best Measure: Reach<sup>1</sup> of Pins deactivated for violating policy

Pinterest does not report on prevalence and instead uses reach as a metric due to the nature of the platform.

Pinterest Policy	Latest Period								Previous Period							
	Q4 2022				Q3 2022				Q2 2022				Q1 2022			
	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people
Medical misinformation	63%	28%	3%	7%	64%	31%	2%	3%	91%	8%	0.3%	0.4%	85%	13%	0.7%	1%
Self-injury and harmful behavior	74%	21%	4%	1%	77%	21%	2%	0.6%	93%	6%	0.6%	0.4%	71%	23%	3%	3%
Spam	33%	49%	18%	0.7%	76%	20%	3%	1%	89%	10%	0.7%	0.3%	60%	34%	4%	1%
Trademark <sup>3</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Violent Actors	63%	28%	6%	3%	73%	21%	4%	2%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

<sup>1</sup> Calculated by looking at each policy-violating Pin deactivated in a reporting period, then counting the number of unique users that saw each of those Pins during the reporting period for at least 1 second, before it was deactivated.

<sup>2</sup> CSE includes any content that might exploit or endanger minors. By sharing reach for CSE content, we are not implying in any way that harm to children is somehow lessened if fewer people see it. The content is violative and wrong, no matter how many people see it. We share the data only to be transparent in our efforts to remove CSE from our platform.

<sup>3</sup> We do not currently report on reach for Trademark.

<sup>4</sup> Reach for Hateful activities and Climate misinformation in Q4 2022 does not include Pins deactivated that we later determined to be false positives and subsequently reinstated. See commentary for more details.





### Question 3: How effective is the platform in enforcing safety policy?

**Authorized Metric:** Reach<sup>1</sup> of Pins deactivated for violating policy

Pinterest Policy	Latest Period								Previous Period							
	Q4 2022				Q3 2022				Q2 2022				Q1 2022			
	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people
Adult content	54%	36%	7%	3%	71%	24%	4%	2%	77%	19%	3%	1%	78%	18%	3%	1%
Adult sexual services	62%	28%	5%	5%	64%	29%	5%	2%	82%	14%	3%	0.9%	95%	5%	0.5%	0.1%
Child sexual exploitation (CSE) <sup>2</sup>	51%	35%	9%	4%	61%	29%	6%	3%	61%	30%	6%	3%	63%	28%	6%	3%
Civic misinformation	82%	16%	1%	1%	64%	30%	4%	2%	46%	47%	5%	3%	20%	67%	5%	8%
Climate Misinformation <sup>4</sup>	1%	72%	8%	19%	27%	60%	7%	6%	13%	81%	4%	2%	N/A	N/A	N/A	N/A
Conspiracy theories	71%	23%	2%	4%	76%	19%	2%	2%	84%	14%	1%	1%	57%	33%	5%	5%
Copyright <sup>3</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Dangerous goods and activities	89%	10%	0.6%	0.3%	85%	13%	0.8%	0.4%	64%	31%	3%	2%	72%	21%	4%	3%
Graphic violence and threats	58%	40%	0.5%	1%	48%	48%	1%	3%	67%	28%	2%	4%	65%	29%	4%	3%
Harassment and criticism	65%	34%	0.8%	0.7%	68%	27%	2%	2%	58%	35%	4%	2%	74%	21%	3%	3%
Hateful activities <sup>4</sup>	52%	35%	4%	9%	52%	37%	5%	6%	62%	26%	4%	8%	63%	22%	6%	9%

<sup>1</sup> Calculated by looking at each policy-violating Pin deactivated in a reporting period, then counting the number of unique users that saw each of those Pins during the reporting period for at least 1 second, before it was deactivated.

<sup>2</sup> CSE includes any content that might exploit or endanger minors. By sharing reach for CSE content, we are not implying in any way that harm to children is somehow lessened if fewer people see it. The content is violative and wrong, no matter how many people see it. We share the data only to be transparent in our efforts to remove CSE from our platform.

<sup>3</sup> We do not currently report on reach for Copyright.

<sup>4</sup> Reach for Hateful activities and Climate misinformation in Q4 2022 does not include Pins deactivated that we later determined to be false positives and subsequently reinstated. See commentary for more details.



### Question 3: How effective is the platform in enforcing safety policy?

**Authorized Metric:** Reach<sup>1</sup> of Pins deactivated for violating policy

Pinterest Policy	Latest Period								Previous Period							
	Q4 2022				Q3 2022				Q2 2022				Q1 2022			
	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people	Seen by 0 people	Seen by 1-9 people	Seen by 10-100 people	Seen by >100 people
Medical misinformation	63%	28%	3%	7%	64%	31%	2%	3%	91%	8%	0.3%	0.4%	85%	13%	0.7%	1%
Self-injury and harmful behavior	74%	21%	4%	1%	77%	21%	2%	0.6%	93%	6%	0.6%	0.4%	71%	23%	3%	3%
Spam	33%	49%	18%	0.7%	76%	20%	3%	1%	89%	10%	0.7%	0.3%	60%	34%	4%	1%
Trademark <sup>3</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Violent Actors	63%	28%	6%	3%	73%	21%	4%	2%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

<sup>1</sup> Calculated by looking at each policy-violating Pin deactivated in a reporting period, then counting the number of unique users that saw each of those Pins during the reporting period for at least 1 second, before it was deactivated.

<sup>2</sup> CSE includes any content that might exploit or endanger minors. By sharing reach for CSE content, we are not implying in any way that harm to children is somehow lessened if fewer people see it. The content is violative and wrong, no matter how many people see it. We share the data only to be transparent in our efforts to remove CSE from our platform.

<sup>3</sup> We do not currently report on reach for Trademark.

<sup>4</sup> Reach for Hateful activities and Climate misinformation in Q4 2022 does not include Pins deactivated that we later determined to be false positives and subsequently reinstated. See commentary for more details.



### Question 3: How effective is the platform in enforcing safety policy?

**Authorized Metric:** Distinct images deactivated<sup>1</sup>, Pins deactivated<sup>2</sup>, Boards deactivated<sup>3</sup>, Accounts deactivated<sup>4</sup>

Violating content deactivated by Pinterest

Pinterest Policy	Latest Period				Q3 2022				Q2 2022				Previous Period			
	Q4 2022				Q3 2022				Q2 2022				Q1 2022			
	Distinct images deactivated	Pins deactivated	Boards deactivated	Accounts deactivated	Distinct images deactivated	Pins deactivated	Boards deactivated	Accounts deactivated	Distinct images deactivated	Pins deactivated	Boards deactivated	Accounts deactivated	Distinct images deactivated	Pins deactivated	Boards deactivated	Accounts deactivated
Adult content	958,673	43,518,977	48,633	4,864	896,950	26,077,118	49,608	7,811	1,058,729	27,606,227	43,757	6,571	975,774	22,309,237	47,965	7,699
Adult sexual services	6,801	105,667	65	121	22,433	212,171	164	157	21,746	51,936	146	156	6,052	208,296	129	117
Child sexual exploitation (CSE) <sup>5</sup>	12,733	1,716,192	1,108	33,228	10,772	687,825	633	21,033	9,085	712,295	1,162	37,694	2,499	300,003	492	10,743
Civic misinformation	3,361	5,488	252	5	1,940	3,778	111	0	1,541	2,782	33	9	712	840	42	10
Climate Misinformation	241	33,040	54	5	164	517	19	0	1,120	1,218	132	7	N/A	N/A	N/A	N/A
Conspiracy theories	1,581	5,988	502	10	2,222	6,870	210	5	3,857	14,312	311	19	2,579	5,039	223	19
Copyright	76,745	17,108,646	0	432	37,835	11,137,479	0	372	41,667	12,898,085	0	381	63,817	13,995,413	0	380
Dangerous goods and activities	10,283	1,657,667	501	531	10,029	2,086,885	642	207	12,253	81,286	583	245	11,308	55,985	603	222
Graphic violence and threats	11,383	839,175	732	41	8,888	224,161	638	36	12,845	96,854	633	60	9,662	124,008	800	120
Harassment and criticism	4,287	618,977	931	236	5,544	136,806	919	169	9,672	232,858	1,356	266	3,128	61,237	931	197
Hateful activities	9,872	196,464	838	95	7,044	80,731	680	78	5,774	40,455	705	76	4,016	27,585	726	105

<sup>1</sup> Does not include distinct images that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.

<sup>2</sup> Does not include Pins that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.

<sup>3</sup> When policy-violating boards are deactivated, all Pins on those boards are removed as well. Does not include boards that were removed because they belonged to a user that was deactivated.

<sup>4</sup> When policy-violating accounts are removed, all Pins and boards belonging to those accounts are removed as well.

<sup>5</sup> CSE includes any content that might exploit or endanger minors. We count all deactivations for CSE, no matter what other actions may have already been taken against the Pin, board or user. For example, if a Pin has been automatically deactivated—meaning no one on the platform can see it—for violating our Spam policy but we later learn that it contains material that violates our CSE policy, the Pin is counted in both our Spam and CSE deactivation numbers.



### Question 3: How effective is the platform in enforcing safety policy?

**Authorized Metric:** Distinct images deactivated<sup>1</sup>, Pins deactivated<sup>2</sup>, Boards deactivated<sup>3</sup>, Accounts deactivated<sup>4</sup>

Violating content deactivated by Pinterest

Pinterest Policy	Latest Period				Previous Period											
	Q4 2022				Q3 2022				Q2 2022				Q1 2022			
	Distinct images deactivated	Pins deactivated	Boards deactivated	Accounts deactivated	Distinct images deactivated	Pins deactivated	Boards deactivated	Accounts deactivated	Distinct images deactivated	Pins deactivated	Boards deactivated	Accounts deactivated	Distinct images deactivated	Pins deactivated	Boards deactivated	Accounts deactivated
Medical misinformation	2,686	19,188	155	6	4,995	42,455	386	2	3,906	262,909	299	4	4,426	113,195	513	17
Self-injury and harmful behavior	88,596	15,166,148	6,535	1,777	57,731	17,052,987	13,529	197	8,494	1,232,828	14,667	29	6,574	107,059	16,144	61
Spam	689,081	739,176	0	2,108,189	47,099	114,450	0	2,248,045	37,398	259,996	10,543	1,680,646	35,243	74,941	0	1,477,695
Trademark	29,807	38,677	107	282	26,688	35,301	339	279	27,125	34,366	164	241	27,859	36,215	241	205
Violent Actors	4,120	257,149	697	95	817	88,345	362	42	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

<sup>1</sup> Does not include distinct images that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.

<sup>2</sup> Does not include Pins that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.

<sup>3</sup> When policy-violating boards are deactivated, all Pins on those boards are removed as well. Does not include boards that were removed because they belonged to a user that was deactivated.

<sup>4</sup> When policy-violating accounts are removed, all Pins and boards belonging to those accounts are removed as well.

<sup>5</sup> CSE includes any content that might exploit or endanger minors. We count all deactivations for CSE, no matter what other actions may have already been taken against the Pin, board or user. For example, if a Pin has been automatically deactivated—meaning no one on the platform can see it—for violating our Spam policy but we later learn that it contains material that violates our CSE policy, the Pin is counted in both our Spam and CSE deactivation numbers.



### Question 3: How effective is the platform in enforcing safety policy?

**Authorized Metric:** Distinct images deactivated<sup>1</sup>, Pins deactivated<sup>2</sup>, Boards deactivated<sup>3</sup>, Accounts deactivated<sup>4</sup>

#### Violating content deactivated by Pinterest

Pinterest policy	Commentary
Adult content	We deactivated fewer Pins for violating this policy in Q3 2022 than in Q4 2022. Of the Pins we deactivated in Q3, 98% were seen by 100 or fewer users in that reporting period.
Adult sexual services	We deactivated more Pins for violating our adult sexual services policy in Q3 2022 than in Q4 2022. Of the Pins we deactivated in Q3, 98% were seen by 100 or fewer users in that reporting period.
Child sexual exploitation (CSE)	<p>Pinterest does not tolerate child sexual exploitation (CSE) of any kind on our platform. That means we enforce a strict, zero-tolerance policy for any content—including imagery, video, or text—that might exploit or endanger minors. Detecting and removing this type of content is of the utmost importance to our Trust and Safety team, and we are proud of our broad-reaching policies and robust efforts to keep our users safe.</p> <p>Pinterest's CSE policy prohibits not just illegal child sexual abuse material, but goes a step further to prohibit any content that contributes to the sexualization of minors.</p>
Civic misinformation	Fighting misinformation is complex and always evolving. Content enforcement numbers may change quarter-to-quarter depending on real-world events. Of the Pins we deactivated in Q4, 98% were seen by fewer than 10 users in that reporting period.
Climate Misinformation	<p>In April 2022, Pinterest launched a new climate misinformation policy to keep false and misleading claims around climate change off the platform. Under our climate misinformation policy, we remove content that denies the existence or impacts of climate change and false or misleading content about natural disasters and extreme weather events. Our climate misinformation policy is yet another step in Pinterest's journey to combat misinformation and create a safe space online.</p> <p>We determined that one of the distinct images deactivated in Q4 2022 was incorrectly deactivated, as were more than 32,000 Pins that matched that image. We reinstated them after identifying the error. Of the Pins that we believe were correctly deactivated, 81% were seen by 100 or fewer users in this reporting period.</p> <p>We've included those false positives in the Q4 enforcement data, but we excluded them from the reach metric for this policy in an effort to provide more accurate insight into the number of users who saw a Pin that actually violates this policy before the Pin was deactivated.</p>
Conspiracy theories	We deactivated fewer Pins for violating this policy in Q4 2022 than in Q3 2022. Of the Pins we deactivated in Q4, 96% were seen by 100 or fewer users in that reporting period.
Copyright	<p>Pinterest has always been a place for content creators, brands and publishers worldwide to feature their content and build value. Many creators upload their own content or encourage users to do so using buttons on their websites designed to facilitate saving to Pinterest and welcome the exposure and user traffic generated when users save images. We work hard to give creators control over their content, including by designating which websites should be linked to and receive traffic from saved images, using features like our "No Pin" code if they wish to restrict saving from their websites.</p> <p>In cases where rights holders do not want their content to appear on Pinterest, we offer multiple copyright reporting mechanisms for content removal. Once we've assessed a copyright notice, we take appropriate action, which may include removing the reported content from Pinterest.</p>
Dangerous goods and activities	We deactivated more Pins for violating this policy in Q3 2022 than in Q4 2022. Of the Pins we deactivated in Q4, more than 99% were seen by fewer than 10 users in that reporting period.

<sup>1</sup> Does not include distinct images that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.

<sup>2</sup> Does not include Pins that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.

<sup>3</sup> When policy-violating boards are deactivated, all Pins on those boards are removed as well. Does not include boards that were removed because they belonged to a user that was deactivated.

<sup>4</sup> When policy-violating accounts are removed, all Pins and boards belonging to those accounts are removed as well.



### Question 3: How effective is the platform in enforcing safety policy?

**Authorized Metric:** Distinct images deactivated<sup>1</sup>, Pins deactivated<sup>2</sup>, Boards deactivated<sup>3</sup>, Accounts deactivated<sup>4</sup>

#### Violating content deactivated by Pinterest

Pinterest policy	Commentary
Graphic violence and threats	We deactivated fewer Pins for violating this policy in Q3 2022 than in Q4 2022. Of the Pins we deactivated in Q4 2022, 99% were seen by 100 or fewer users in that reporting period.
Harassment and criticism	We deactivated fewer Pins for violating this policy in Q3 2022 than in Q4 2022. Of the Pins we deactivated in Q4, more than 99% were seen by 100 or fewer users in that reporting period.
Hateful activities	We determined that two of the distinct images deactivated in Q4 2022 were incorrectly deactivated, as were more than 100,000 Pins that matched those images. We reinstated them after identifying the error. Of the Pins that we believe were correctly deactivated, 91% were seen by 100 or fewer users in this reporting period.  We've included those false positives in the Q4 enforcement data, but we excluded them from the reach metric for this policy in an effort to provide more accurate insight into the number of users who saw a Pin that actually violates this policy before the Pin was deactivated.
Medical misinformation	Pinterest is deeply committed to combating health misinformation. We continue to engage with public health experts to stay on top of trends and get feedback on our policy and enforcement approaches for topics such as medical misinformation.  We deactivated more Pins for violating this policy in Q3 2022 than in Q4 2022. Of the Pins we deactivated in Q3, more than 97% were seen by 100 or fewer users in that reporting period.
Self-injury and harmful behavior	We continued investing in work to improve content moderation for self-harm content. As a result, we saw a large increase in Pins deactivated throughout Q3 and Q4. Of the Pins we deactivated in Q3, more than 99% were seen by 100 or fewer users in that reporting period.
Spam	We use the latest in machine learning technology to build automated models that swiftly detect and act against spam of all kinds. We iterate on these models at regular intervals by adding new data and exploring new technical breakthroughs to either maintain or improve their performance over time to effectively address spam. Given the adversarial, iterative nature of fighting spam, content enforcement numbers may change quarter-to-quarter, especially after a large attack.  We deactivated fewer Pins for violating this policy in Q3 2022 than in Q4 2022. Of the Pins we deactivated in Q4, more than 99% were seen by 100 or fewer users in that reporting period.
Trademark	Pinterest respects the trademark rights of others. Trademark owners can contact us through our reporting mechanisms if they have concerns that someone may be using their trademark in an infringing way on our site. We review submissions we receive and take appropriate action, including removal of the content from Pinterest.
Violent Actors	Pinterest isn't a place for violent content, groups or individuals. We limit the distribution of or remove content and accounts that encourage, praise, promote, or provide aid to dangerous actors or groups and their activities. This includes extremists, terrorist organizations, and gangs and other criminal organizations. We work with industry, government and security experts to help us identify these groups.

<sup>1</sup> Does not include distinct images that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.

<sup>2</sup> Does not include Pins that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.

<sup>3</sup> When policy-violating boards are deactivated, all Pins on those boards are removed as well. Does not include boards that were removed because they belonged to a user that was deactivated.

<sup>4</sup> When policy-violating accounts are removed, all Pins and boards belonging to those accounts are removed as well.



### Question 3: How effective is the platform in enforcing safety policy?

**Authorized Metric:** How Pins are deactivated

Percentage of violating Pins deactivated by enforcement mechanism

Pinterest Policy	Latest Period						Previous Period					
	Q4 2022			Q3 2022			Q2 2022			Q1 2022		
	Automated <sup>1</sup>	Manual <sup>2</sup>	Hybrid <sup>3</sup>	Automated <sup>1</sup>	Manual <sup>2</sup>	Hybrid <sup>3</sup>	Automated <sup>1</sup>	Manual <sup>2</sup>	Hybrid <sup>3</sup>	Automated <sup>1</sup>	Manual <sup>2</sup>	Hybrid <sup>3</sup>
Adult content	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%
Adult sexual services	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%
Child sexual exploitation (CSE)	0%	<1%	>99%	0%	2%	98%	0%	1%	99%	0%	<1%	>99%
Civic misinformation	0%	88%	12%	0%	86%	14%	0%	69%	31%	0%	90%	10%
Climate Misinformation	0%	1%	99%	N/A	33%	67%	0%	>99%	<1%	N/A	N/A	N/A
Conspiracy theories	0%	41%	59%	0%	45%	55%	0%	77%	23%	0%	67%	33%
Copyright	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%
Dangerous goods and activities	<1%	<1%	>99%	<1%	<1%	>99%	7%	12%	81%	12%	16%	72%
Graphic violence and threats	0%	2%	98%	0%	5%	95%	0%	27%	73%	<1%	9%	91%
Harassment and criticism	0%	<1%	>99%	0%	5%	95%	0%	4%	96%	0%	5%	95%
Hateful activities	0%	7%	93%	0%	11%	89%	0%	15%	85%	0%	16%	84%

<sup>1</sup> Our automated tools use a combination of signals to identify and take action against potentially violating content. Our machine learning models assign scores to each image added to our platform. Using these scores, our automated tools can then apply the same enforcement decision to other Pins containing the same image.

<sup>2</sup> We manually deactivate Pins through our human review process. Pins deactivated through this process may include those identified internally and those reported to us by third parties. It also includes the Pins that are reviewed and deactivated by one of our team members after a user report.

<sup>3</sup> Hybrid deactivations include those where a human determines that a Pin violates policy, and automated systems expand that decision to enforce against machine-identified matching Pins. Depending on the volume of matching Pins, a hybrid deactivation may result in a number of Pins deactivated or none at all.



### Question 3: How effective is the platform in enforcing safety policy?

**Authorized Metric:** How Pins are deactivated

Percentage of violating Pins deactivated by enforcement mechanism

Pinterest Policy	Latest Period						Previous Period					
	Q4 2022			Q3 2022			Q2 2022			Q1 2022		
	Automated <sup>1</sup>	Manual <sup>2</sup>	Hybrid <sup>3</sup>	Automated <sup>1</sup>	Manual <sup>2</sup>	Hybrid <sup>3</sup>	Automated <sup>1</sup>	Manual <sup>2</sup>	Hybrid <sup>3</sup>	Automated <sup>1</sup>	Manual <sup>2</sup>	Hybrid <sup>3</sup>
Medical misinformation	22%	7%	72%	14%	9%	77%	2%	2%	97%	4%	3%	93%
Self-injury and harmful behavior	3%	<1%	97%	4%	<1%	96%	0%	<1%	>99%	0%	6%	94%
Spam	100%	0%	0%	100%	0%	0%	>99%	<1%	0%	>99%	<1%	<1%
Trademark	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%
Violent Actors	0%	<1%	>99%	0%	<1%	>99%	N/A	N/A	N/A	N/A	N/A	N/A

<sup>1</sup> Our automated tools use a combination of signals to identify and take action against potentially violating content. Our machine learning models assign scores to each image added to our platform. Using these scores, our automated tools can then apply the same enforcement decision to other Pins containing the same image.

<sup>2</sup> We manually deactivate Pins through our human review process. Pins deactivated through this process may include those identified internally and those reported to us by third parties. It also includes the Pins that are reviewed and deactivated by one of our team members after a user report.

<sup>3</sup> Hybrid deactivations include those where a human determines that a Pin violates policy, and automated systems expand that decision to enforce against machine-identified matching Pins. Depending on the volume of matching Pins, a hybrid deactivation may result in a number of Pins deactivated or none at all.





## Question 4: How does the platform perform at correcting mistakes?

**Authorized Metric:** Account Appeals, Account Reinstatements

Accounts appealed after a deactivation, Accounts reinstated after an appeal

Pinterest Policy	Latest Period				Previous Period			
	Q4 2022		Q3 2022		Q2 2022		Q1 2022	
	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated
Adult content	1,539	730	2,049	1,290	1,674	1,070	2,104	1,229
Adult sexual services	21	2	4	1	4	1	9	3
Child sexual exploitation (CSE)	5,731	2,686	3,896	2,053	7,467	5,971	2,164	1,169
Civic misinformation	5	2	2	0	4	2	4	1
Climate Misinformation	3	3	0	0	0	0	N/A	N/A
Conspiracy theories	14	4	5	2	37	8	6	4
Copyright <sup>1</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Dangerous goods and activities	21	1	14	2	11	3	12	2
Graphic violence and threats	24	8	13	4	25	14	25	9
Harassment and criticism	93	57	50	37	55	38	28	16
Hateful activities	40	12	23	6	33	12	31	6

<sup>1</sup> We do not currently report on account appeals and reinstatements for Copyright or Trademark.



## Question 4: How does the platform perform at correcting mistakes?

**Authorized Metric:** Account Appeals, Account Reinstatements

Accounts appealed after a deactivation, Accounts reinstated after an appeal

Pinterest Policy	Latest Period				Previous Period			
	Q4 2022		Q3 2022		Q2 2022		Q1 2022	
	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated
Medical misinformation	5	3	2	1	1	1	4	1
Self-injury and harmful behavior	199	44	31	12	21	10	30	9
Spam	76,049	56,137	134,405	109,381	73,639	55,930	98,156	79,469
Trademark <sup>1</sup>	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Violent Actors	8	3	1	1	N/A	N/A	N/A	N/A

<sup>1</sup> We do not currently report on account appeals and reinstatements for Copyright or Trademark.



### Safety by Design: Our Commitment to Users & Advertisers

Safety is central to Snap’s mission. It is a principle that drives many of our unique design and policy choices, such as opening up to the camera rather than a feed of unmoderated content. Across Snapchat, we have built safety into the fundamental architecture of our platform.

Instead of developing an open news feed that optimizes for unchecked, sensational, or toxic content, we built Discover—a content platform for vetted publishers and creators with a proven audience. Like a television network, our editorial teams choose the publishers and shows that run on Discover. We feature high profile, respected partners, like Disney, NBCU, The Washington Post, Bloomberg, NFL, and NBA, mainstream celebrities like Meghan Trainor and Jack Harlow, and popular creators like David Dobrik.

In order to participate on Discover, our editorial partners must comply with our content guidelines—which prohibit the spread of misinformation, hate speech, conspiracy theories, violence and many other categories of harmful content.

### Machine Learning + Human Review = Better Content Moderation

We have always taken a forward-leaning approach to content moderation and enforcing our policies. We combine machine learning with a dedicated team of real people to detect and review potentially inappropriate content in public posts.

For example, on Spotlight, where creators can submit creative and entertaining videos to share with the broader Snapchat community, all content is first reviewed automatically by artificial intelligence before being widely distributed. Once a piece of content gains more viewership, it’s then reviewed by human moderators before it is given the opportunity to reach a large audience.

### Updates to Transparency Reports

As part of our ongoing commitment toward improving our transparency reports, we introduced several new elements to the latest edition.

We are now providing increased insight into our efforts to combat Child Sexual Exploitation and Abuse Imagery (CSEAI). Moving forward, we will be sharing insight on total CSEAI content that we enforced against by removing, as well as the total number of CSEAI reports\* (i.e., “CyberTips”) that we made to the U.S. National Center for Missing and Exploited Children (NCMEC).

For the first time, we are introducing “False Information” as a stand-alone category available at the country level, building on our previous practice of reporting false information globally.

When looking at the data in the report, note that the figures for total reports and enforcement only count the content which is reported to us. It does not count the instances where Snap proactively detected and took action against content before it was reported. We believe that the improvements we’ve made to our proactive detection efforts played a large role in the decrease of total reports, enforcement numbers, and turnaround times.

Because our enhanced, automated-detection tools often identified and removed content before it had a chance to reach Snapchatters, we saw a decrease in reactive content enforcements (i.e., reports from Snapchatters). Since our last report, we saw a 44% decrease in threatening and violent content enforcements on reports from Snapchatters, as well as a 37% decrease in drug content enforcements and a 34% decrease in hate speech content enforcements. On average, our median turnaround time for removing violating content has improved 33%.

While Snapchat has evolved over the years, our commitment to transparency and prioritizing the safety and well-being of our community remains unchanged.



# Question 1: How safe is the platform for consumers?

## Authorized Metric: Violative view rate

An estimate of the percentage of Snap and Story views that violated our community guidelines in a given reporting period.

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
<b>Adult &amp; Explicit Sexual Content</b>	Sexual Content	During the reporting period, we saw a Violative View Rate (VVR) of 0.04 percent	During the reporting period, we saw a Violative View Rate (VVR) of 0.04 percent	We prohibit promoting, distributing, or sharing pornographic content. We also don't allow commercial activities that relate to pornography or sexual interactions (whether online or offline). Breastfeeding and other depictions of nudity in non-sexual contexts are generally permitted. We prohibit any activity that involves sexual exploitation or abuse of a minor, including sharing child sexual exploitation or abuse imagery, grooming, or sexual extortion (sextortion). We report all instances of child sexual exploitation to authorities, including attempts to engage in such conduct. Never post, save, send, forward, distribute, or ask for nude or sexually explicit content involving anyone under the age of 18 (this includes sending or saving such images of yourself).
<b>Arms &amp; Ammunition</b>	Illegal or Regulated Activities			Don't use Snapchat for any illegal activity. This includes promoting, facilitating, or participating in criminal activity, such as buying, selling, exchanging, or facilitating sales of illegal or regulated drugs, contraband (such as child sexual abuse or exploitation imagery), weapons, or counterfeit goods or documents. It also includes promoting or facilitating any form of exploitation, including human trafficking or sex trafficking.
<b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b>	Threats, Violence & Harm			Encouraging or engaging in violent or dangerous behavior is prohibited. Never intimidate or threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence, including animal abuse, are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury, suicide or eating disorders.
<b>Death, Injury or Military Conflict</b>	Threats, Violence & Harm			Encouraging or engaging in violent or dangerous behavior is prohibited. Never intimidate or threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence, including animal abuse, are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury, suicide or eating disorders.
<b>Online piracy</b>	Harmful False or Deceptive Information			We prohibit pretending to be someone (or something) that you're not, or attempting to deceive people about who you are. This includes impersonating your friends, celebrities, public figures, brands, or other organizations for harmful, non-satirical purposes. We prohibit spam and deceptive practices, including imitating Snapchat or Snap Inc.
<b>Misinformation</b>	Harmful False or Deceptive Information			We prohibit spreading false information that causes harm or is malicious, such as denying the existence of tragic events, unsubstantiated medical claims, undermining the integrity of civic processes, or manipulating content for false or misleading purposes.





# Question 1: How safe is the platform for consumers?

## Authorized Metric: Violative view rate

An estimate of the percentage of Snap and Story views that violated our community guidelines in a given reporting period.

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
<b>Hate speech &amp; acts of aggression</b>	Hateful Content, Terrorism, and Violent Extremism	During the reporting period, we saw a Violative View Rate (VVR) of 0.04 percent	During the reporting period, we saw a Violative View Rate (VVR) of 0.04 percent	Hate speech or content that demeans, defames, or promotes discrimination or violence on the basis of race, color, caste, ethnicity, national origin, religion, sexual orientation, gender identity, disability, or veteran status, immigration status, socio-economic status, age, weight, or pregnancy status is prohibited.
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b>	Threats, Violence & Harm			As standalone, obscenity and profanity do not constitute a violation of Snap's Community Guidelines. If categorized separately (e.g. profanity that is also hate speech), takedown would be reported in the appropriate, corresponding category. Additionally, Snaps of gratuitous or graphic violence, including animal abuse, are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury, suicide or eating disorders.
<b>Illegal Drugs / Tobacco / e-cigarettes / Vaping / Alcohol</b>	Illegal or Regulated Activities			We prohibit the illegal promotion of regulated goods or industries, including unauthorized promotion of gambling, tobacco or vape products, and alcohol.
<b>Spam or Harmful Content</b>	Harmful False or Deceptive Information			We prohibit spam and other deceptive practices, including manipulating content for misleading purposes or to imitate Snap Inc. or Snapchat.
<b>Terrorism</b>	Hateful Content, Terrorism, and Violent Extremism			Terrorist organizations, violent extremists, and hate groups are prohibited from using our platform. We have no tolerance for content that advocates or advances terrorism or violent extremism.
<b>Debated Sensitive Social Issue</b>	Harmful, False, or Deceptive Information			We prohibit spreading false information that causes harm or is malicious, such as denying the existence of tragic events, unsubstantiated medical claims, undermining the integrity of civic processes, or manipulating content for false or misleading purposes.





## Question 2: How safe is the platform for advertisers?

### Authorized Metric: Violative view rate

An estimate of the percentage of Snap and Story views that violated our community guidelines in a given reporting period.

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
<b>Adult &amp; Explicit Sexual Content</b>	Sexual Content	During the reporting period, we saw a Violative View Rate (VVR) of 0.04 percent	During the reporting period, we saw a Violative View Rate (VVR) of 0.04 percent	We prohibit promoting, distributing, or sharing pornographic content. We also don't allow commercial activities that relate to pornography or sexual interactions (whether online or offline). Breastfeeding and other depictions of nudity in non-sexual contexts are generally permitted. We prohibit any activity that involves sexual exploitation or abuse of a minor, including sharing child sexual exploitation or abuse imagery, grooming, or sexual extortion (sextortion). We report all instances of child sexual exploitation to authorities, including attempts to engage in such conduct. Never post, save, send, forward, distribute, or ask for nude or sexually explicit content involving anyone under the age of 18 (this includes sending or saving such images of yourself).
<b>Arms &amp; Ammunition</b>	Illegal or Regulated Activities			Don't use Snapchat for any illegal activity. This includes promoting, facilitating, or participating in criminal activity, such as buying, selling, exchanging, or facilitating sales of illegal or regulated drugs, contraband (such as child sexual abuse or exploitation imagery), weapons, or counterfeit goods or documents. It also includes promoting or facilitating any form of exploitation, including human trafficking or sex trafficking.
<b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b>	Threats, Violence & Harm			Encouraging or engaging in violent or dangerous behavior is prohibited. Never intimidate or threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence, including animal abuse, are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury, suicide or eating disorders.
<b>Death, Injury or Military Conflict</b>	Threats, Violence & Harm			Encouraging or engaging in violent or dangerous behavior is prohibited. Never intimidate or threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence, including animal abuse, are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury, suicide or eating disorders.
<b>Online piracy</b>	Harmful False or Deceptive Information			We prohibit pretending to be someone (or something) that you're not, or attempting to deceive people about who you are. This includes impersonating your friends, celebrities, public figures, brands, or other organizations for harmful, non-satirical purposes. We prohibit spam and deceptive practices, including imitating Snapchat or Snap Inc.
<b>Misinformation</b>	Harmful False or Deceptive Information			We prohibit spreading false information that causes harm or is malicious, such as denying the existence of tragic events, unsubstantiated medical claims, undermining the integrity of civic processes, or manipulating content for false or misleading purposes.



## Question 2: How safe is the platform for advertisers?

### Authorized Metric: Violative view rate

An estimate of the percentage of Snap and Story views that violated our community guidelines in a given reporting period.

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
<b>Hate speech &amp; acts of aggression</b>	Hateful Content, Terrorism, and Violent Extremism	During the reporting period, we saw a Violative View Rate (VVR) of 0.04 percent	During the reporting period, we saw a Violative View Rate (VVR) of 0.04 percent	Hate speech or content that demeans, defames, or promotes discrimination or violence on the basis of race, color, caste, ethnicity, national origin, religion, sexual orientation, gender identity, disability, or veteran status, immigration status, socio-economic status, age, weight, or pregnancy status is prohibited.
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b>	Threats, Violence & Harm			As standalone, obscenity and profanity do not constitute a violation of Snap's Community Guidelines. If categorized separately (e.g. profanity that is also hate speech), takedown would be reported in the appropriate, corresponding category. Additionally, Snaps of gratuitous or graphic violence, including animal abuse, are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury, suicide or eating disorders.
<b>Illegal Drugs / Tobacco / e-cigarettes / Vaping / Alcohol</b>	Illegal or Regulated Activities			We prohibit the illegal promotion of regulated goods or industries, including unauthorized promotion of gambling, tobacco or vape products, and alcohol.
<b>Spam or Harmful Content</b>	Harmful False or Deceptive Information			We prohibit spam and other deceptive practices, including manipulating content for misleading purposes or to imitate Snap Inc. or Snapchat.
<b>Terrorism</b>	Hateful Content, Terrorism, and Violent Extremism			Terrorist organizations, violent extremists, and hate groups are prohibited from using our platform. We have no tolerance for content that advocates or advances terrorism or violent extremism.
<b>Debated Sensitive Social Issue</b>	Harmful, False, or Deceptive Information			We prohibit spreading false information that causes harm or is malicious, such as denying the existence of tragic events, unsubstantiated medical claims, undermining the integrity of civic processes, or manipulating content for false or misleading purposes.



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

#### Authorized Metric: Removals of Violating Content, Removals of Violating Accounts

Content removed by Snap - Users removed by Snap

Some individual Snap categories encompass multiple GARM categories, e.g. GARM's "Crime & Harmful acts to individuals and Society, Human Right Violations" and "Death, Injury or Military Conflict" categories both roll up under "Threats, Violence & Harm" in Snap's Transparency Report. Depending on report consolidation methodologies, calling this out to ensure that actioned accounts and content aren't inadvertently double counted because some are listed twice in this response.

GARM Category	Relevant Snap Policy	Violation Reason	Latest Period		Previous Period		Commentary
			Content Actioned	Actors Actioned	Content Actioned	Actors Actioned	
<b>Adult &amp; Explicit Sexual Content</b>	Sexual Content	Sexually Explicit Content	4,355,143	1,972,482	4,869,272	1,716,547	We prohibit promoting, distributing, or sharing pornographic content. We also don't allow commercial activities that relate to pornography or sexual interactions (whether online or offline). Breastfeeding and other depictions of nudity in non-sexual contexts are generally permitted. We prohibit any activity that involves sexual exploitation or abuse of a minor, including sharing child sexual exploitation or abuse imagery, grooming, or sexual extortion (sextortion). We report all instances of child sexual exploitation to authorities, including attempts to engage in such conduct. Never post, save, send, forward, distribute, or ask for nude or sexually explicit content involving anyone under the age of 18 (this includes sending or saving such images of yourself).
<b>Arms &amp; Ammunition</b>	Illegal or Regulated Activities	Weapons	27,722	23,181	28,706	21,310	Don't use Snapchat for any illegal activity. This includes promoting, facilitating, or participating in criminal activity, such as buying, selling, exchanging, or facilitating sales of illegal or regulated drugs, contraband (such as child sexual abuse or exploitation imagery), weapons, or counterfeit goods or documents. It also includes promoting or facilitating any form of exploitation, including human trafficking or sex trafficking.
<b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b>	Threats, Violence & Harm	Threats & Violence	150,199	113,887	232,565	159,214	Encouraging or engaging in violent or dangerous behavior is prohibited. Never intimidate or threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence, including animal abuse, are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury, suicide or eating disorders.
<b>Death, Injury or Military Conflict</b>	Threats, Violence & Harm	Threats & Violence	150,199	113,887	232,565	159,214	Encouraging or engaging in violent or dangerous behavior is prohibited. Never intimidate or threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence, including animal abuse, are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury, suicide or eating disorders.
<b>Online piracy</b>	Harmful, False, or Deceptive Information	Spam	348,790	289,511	153,621	110,102	We prohibit pretending to be someone (or something) that you're not, or attempting to deceive people about who you are. This includes impersonating your friends, celebrities, public figures, brands, or other organizations for harmful, non-satirical purposes. We prohibit spam and deceptive practices, including imitating Snapchat or Snap Inc.
<b>Misinformation</b>	Harmful False or Deceptive Information	False Information	4,877	3,371	-	-	We prohibit spreading false information that causes harm or is malicious, such as denying the existence of tragic events, unsubstantiated medical claims, undermining the integrity of civic processes, or manipulating content for false or misleading purposes.







### Question 3: How Effective is the Platform in Enforcing Safety Policy?

#### Authorized Metric: Removals of Violating Content, Removals of Violating Accounts

Content removed by Snap - Users removed by Snap

Some individual Snap categories encompass multiple GARM categories, e.g. GARM's "Crime & Harmful acts to individuals and Society, Human Right Violations" and "Death, Injury or Military Conflict" categories both roll up under "Threats, Violence & Harm" in Snap's Transparency Report. Depending on report consolidation methodologies, calling this out to ensure that actioned accounts and content aren't inadvertently double counted because some are listed twice in this response.

GARM Category	Relevant Snap Policy	Violation Reason	Latest Period		Previous Period		Commentary
			Content Actioned	Actors Actioned	Content Actioned	Actors Actioned	
<b>Hate speech &amp; acts of aggression</b>	Hateful Content, Terrorism, and Violent Extremism	Hate Speech	62,069	52,849	93,341	63,767	Hate speech or content that demeans, defames, or promotes discrimination or violence on the basis of race, color, caste, ethnicity, national origin, religion, sexual orientation, gender identity, disability, or veteran status, immigration status, socio-economic status, age, weight, or pregnancy status is prohibited.
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b>	Threats, Violence & Harm	Threats & Violence	150,199	113,887	232,565	159,214	As standalone, obscenity and profanity do not constitute a violation of Snap's Community Guidelines. If categorized separately (e.g. profanity that is also hate speech), takedown would be reported in the appropriate, corresponding category. Additionally, Snaps of gratuitous or graphic violence, including animal abuse, are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury, suicide or eating disorders.
<b>Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol</b>	Illegal or Regulated Activities	Drugs	270,810	200,022	428,311	278,304	We prohibit the illegal promotion of regulated goods or industries, including unauthorized promotion of gambling, tobacco or vape products, and alcohol.
<b>Spam or Harmful Content</b>	Harmful, False, or Deceptive Information	Spam	348,790	289,511	153,621	110,102	We prohibit spam and other deceptive practices, including manipulating content for misleading purposes or to imitate Snap Inc. or Snapchat.
<b>Terrorism</b>	Hateful Content, Terrorism, and Violent Extremism	Terrorist and Violent Extremist Content	-	73	-	22	Terrorist organizations, violent extremists, and hate groups are prohibited from using our platform. We have no tolerance for content that advocates or advances terrorism or violent extremism.
<b>Debated Sensitive Social Issue</b>	Harmful, False, or Deceptive Information	False Information	4,877	3,371	-	-	We prohibit spreading false information that causes harm or is malicious, such as denying the existence of tragic events, unsubstantiated medical claims, undermining the integrity of civic processes, or manipulating content for false or misleading purposes.



### Question 3: How Effective is the Platform in Enforcing Safety Policy?

**Authorized Metric:** Removals of Violating Content expressed by how many times it has been viewed

GARM Category	Relevant Policy	Latest Period				Previous Period				Commentary
		0	<10	10-100	100+	0	<10	10-100	100+	
Adult & Explicit Sexual Content										Snap does not report on this metric at this time
Arms & Ammunition										
Crime & Harmful acts to individuals and Society, Human Right Violations										
Death, Injury or Military Conflict										
Online piracy										
Hate speech & acts of aggression										
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust										
Illegal Drugs / Tobacco /e-cigarettes / Vaping / Alcohol										
Spam or Harmful Content										
Terrorism										
Debated Sensitive Social Issue										



### Question 4: How does the platform perform at correcting mistakes?

**Authorized Metric:** Appeals or Reinstatements

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Content Appealed	Content Reinstated	Content Appealed	Content Reinstated	
Adult & Explicit Sexual Content						Snap does not report on this metric at this time
Arms & Ammunition						
Crime & Harmful acts to individuals and Society, Human Right Violations						
Death, Injury or Military Conflict						
Online piracy						
Hate speech & acts of aggression						
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust						
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol						
Spam or Harmful Content						
Terrorism						
Debated Sensitive Social Issue						





At Twitch, we strive to create welcoming, interactive spaces where diverse global communities can express themselves safely and feel like they belong. Our goal is to foster an environment that supports and sustains streamers, welcomes and entertains viewers, and minimizes the prevalence of harmful interactions. For Twitch, this means deterring harm while giving streamers the guidelines, tools, technology and education they need to build vibrant communities with their own distinct standards and norms. We also believe every community member is an important and active contributor to a safer Twitch, and work to cultivate an always-on safety dialogue to promote a safe and welcoming culture. Community feedback guides all aspects of our safety journey, from UserVoice, to the Safety Advisory Council and Creator Camps.

Community safety is our top priority, and one of our largest areas of investment. Like our community, safety at Twitch is constantly evolving. Safety is never an “end state,” and we’re always iterating on existing tools and policies, fortifying our proactive detection and operations behind the scenes, and working on new updates to come.

## H2 2022 Overview

At Twitch, we are constantly investing in tooling to ensure all of our streamers and viewers are authentically, safely, and meaningfully engaging with each other.

In H2 2022, we made updates our Community Guidelines to be simpler, easier to understand, and more intuitive to navigate. This was to explain how our guidelines contribute to making Twitch better, and give more clarity overall into our policies and how they apply across Twitch’s global community.

Additionally, we instituted an update which prohibited streaming certain sites that may contain slots, roulette, and dice games and are not licensed in the US or other jurisdictions that offer consumer protections such as deposit limits, waiting periods, and age verification systems.

We’re continuing to expand and improve upon tools we offer to users. We introduced Shared Ban Info which expands upon our existing an Evasion Detection tool (formerly known as Suspicious User Detection). Shared Ban Info targets chat ban evasion and allows streamers to share with one another information about who has been banned from their channels. This allows streamers to collaborate with one another to help keep serial harassers out of their communities. We also introduced Shield Mode which gives streamers and their mods a way to quickly enable additional pre-set safety settings in the face of harassment or an influx of bad actors.

We worked with external organizations on additional education resources, which focused on media literacy and resources for parents. Twitch worked with MediaWise, a non-profit focused on improving digital literacy, to host a livestream and produce [educational videos](#) to help boost personal online media literacy skills.

We also collaborated with ConnectSafely, a non-profit focused on online safety, privacy, and wellness education, to create resources for parents interested in learning more about Twitch. ConnectSafely’s [Parent’s Guide](#) to Twitch is available on their website and gives parents a resource for parents looking for more information about how Twitch works and advice for helping teens stay safe on Twitch.

## Methodology

Twitch is a live-streaming service, thus the vast majority of the content viewed on Twitch is ephemeral. For this reason, we do not consider content removal as the primary means of enforcing adherence to our Community Guidelines. Content is flagged by either machine detection or via user-submitted reports, and our team of experienced specialists is responsible for reviewing these reports and issuing the appropriate enforcements for verified violations.



### Prevalence Metrics

Twitch measures prevalence normalized by violative views. Specifically we use the percentage of Hours Watched (HWs) on content that violates the Twitch Community Guidelines. Our violative view rate metric includes content that does not fall into a [GARM sensitive content category](#), but still violates our guidelines because certain violative content falls within the topics or categories that are not part of the GARM Framework.

We calculate the violative view metric by looking at any enforcement action issued and aggregating hours watched on content that resulted in enforcement. We approximate hours watched by aggregating hours watched on enforced content for the day when the report was filed. Using the same methodology, aggregating impressions delivered on the day when a channel receives a violation, Twitch measures advertising safety error rate as a % of total advertising impressions delivered on content that violates our Twitch Community Guidelines.

### Methodology Limitations

We measure violative content by aggregating content that is reported by our users or flagged by our automated machine detection tools, and issued an enforcement action. This methodology excludes violative content that is not user reported or not flagged by our automated tools and therefore is potentially an undercount. (Violative content with high viewership, that is therefore more impactful on the metric, has a higher likelihood of getting reported.)

For any enforcement action, we consider the timestamp of the user and machine detection reports that resulted in the enforcement and aggregate the HWs or impressions for that day. This approximation has limitations: due to the live and ephemeral nature of the livestream it is possible not all viewership on the channel for that day comes from the violative content and for VOD content, it is possible the violative content is viewed for much longer than a day. VOD content is not the main form of content consumption on Twitch and most users do not view VODs.

We measure violative content as any content that does not meet our Community Guidelines. This is a broader definition than GARM brand safety floor definitions, and we risk overstating the violative HWs / Impressions when viewed against the GARM content categories.

Our Community Guidelines cover similar content as the GARM sensitive content categories; however, due to the differences in how we categorize and define this content, there is some overlap in our reporting for each content category. Where relevant in the GARM Aggregated Measurement report, Twitch included multiple Community Guidelines that fit into each GARM content category.

We chose not to report on two categories - Arms and Ammunition, and Debated Sensitive Issues. Enforcements related to the GARM definition of Arms and Ammunition is dispersed and counted under other categories. While Twitch does not count these enforcements separately, our policies do prohibit use of firearms that may endanger life and align with the GARM safety floor and definition. Many of the violative behaviors under GARM definition in Debated Sensitive Issues are addressed in Twitch's policies of Safety and Civility & Respect.



## Question 1: How safe is the platform for consumers?

### Authorized Metric: Violative View Rate

Twitch measures consumer safety as a % of Hours Watched (HWs) on content that is deemed violative of the Twitch [Community Guidelines](#). This includes content that does not fall into a [GARM sensitive content category](#), but still violates our guidelines.

GARM Category	Latest Period % of total HW	Previous Period % of total HW	Commentary
<b>Adult and Explicit Sexual Content</b>	0.01%	0.01%	We prohibit content that involves nudity, and sexually explicit content. These are standards across live, image, and game content.
<b>Crime and Harmful Acts to Individuals and Society, Human Right Violations</b>	0.01%	<0.01%	We prohibit content and conduct that may be upsetting or damaging, including content that focuses on violence, sexual violence, violent threats, self-harm behaviors, animal cruelty, dangerous or distracted driving, and other illegal, disturbing or frightening content/conduct. For a complete breakdown of our hate and harassment reports and enforcement, please see our <a href="#">Transparency Report</a> .
<b>Death, Injury or Military Conflict</b>	<0.01%	<0.01%	We prohibit content that is upsetting or damaging, including media and conduct that focuses on extreme gore, violence, or violent threats. We may temporarily remove the channel and associated content in situations where a user has lost control of their broadcast due to severe injury, medical emergency, police action, or being targeted with serious violence.
<b>Hate Speech and Acts of Aggression</b>	0.04%	0.01%	We do not tolerate conduct or speech that is hateful or that encourages or incites others to engage in hateful conduct. This includes inciting targeted community abuse, and expressions of hatred based on an identity-based protected characteristic.  While this % of total HW increased by 0.03%, we expect this to fluctuate between 0.01 - 0.05%.
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic, or repulsive content</b>	0.02%	0.02%	We don't permit streamers to be fully or partially nude. Additionally, content that exclusively focuses on extreme or gratuitous gore and violence is prohibited.



## Question 1: How safe is the platform for consumers?

### Authorized Metric: Violative View Rate

Twitch measures consumer safety as a % of Hours Watched (HWs) on content that is deemed violative of the Twitch [Community Guidelines](#). This includes content that does not fall into a [GARM sensitive content category](#), but still violates our guidelines.

GARM Category	Latest Period % of total HW	Previous Period % of total HW	Commentary
<b>Online Piracy</b>	<0.01%	<0.01%	We only allow sharing of content that streamers own, or otherwise have rights to or are authorized to share on Twitch. We do not allow pirated games or content from unauthorized private servers, movies, television shows, or sports matches, music streamers do not own the rights to share, goods or services protected by trademark, or other Twitch streamers' content if the steamer does not have authorization.
<b>Illegal Drugs / Tobacco / e-cigarettes / Vaping / Alcohol</b>	<0.01%	<0.01%	We do not permit any activity that may endanger a streamer's life or lead to physical harm. This includes illegal use of drugs and dangerous consumption of alcohol.
<b>Spam or Harmful Content</b>	0.02%	0.02%	We prohibit disruptive activities such as spamming, because these types of activities violate the integrity of Twitch services, and diminish users' experiences on Twitch.
<b>Terrorism</b>	<0.01%	<0.01%	We do not allow content that depicts, glorifies, encourages, or supports terrorism, or violent extremist actors or acts. This includes threatening to or encouraging others to commit acts that would result in serious physical harm to groups of people or significant property destruction. This metric includes the display or linking of terrorist or extremist propaganda, including graphic pictures or footage of terrorist or extremist violence, even for the purposes of denouncing such content.
<b>Other Violations</b>	0.28%	0.20%	For more information, on content that violates the Twitch guidelines, as well as more detailed takedown rates, please see our <a href="#">Community Guidelines</a> and <a href="#">Transparency Report</a> .



## Question 2: How safe is the platform for advertisers?

### Authorized Metric: Advertising Safety Error Rate

Twitch measures advertising safety error rate as a % of total advertising impressions delivered on content violative of the Twitch [Community Guidelines](#). We use the same methodology as that for violative view rate by aggregating impressions delivered on the day when a channel receives a violation.

GARM Category	Latest Period % of total Impressions	Previous Period % of total Impressions	Commentary
<b>Adult and Explicit Sexual Content</b>	0.02%	0.01%	We prohibit content that involves nudity, and sexually explicit content. These are standards across live, image, and game content.
<b>Crime and Harmful Acts to Individuals and Society, Human Right Violations</b>	0.10%	0.14%	We prohibit content and conduct that may be upsetting or damaging, including content that focuses on violence, sexual violence, violent threats, self-harm behaviors, animal cruelty, dangerous or distracted driving, and other illegal, disturbing or frightening content/conduct. For a complete breakdown of our hate and harassment reports and enforcement, please see our <a href="#">Transparency Report</a> .
<b>Death, Injury or Military Conflict</b>	0.02%	<0.01%	We prohibit content that is upsetting or damaging, including media and conduct that focuses on extreme gore, violence, or violent threats. We may temporarily remove the channel and associated content in situations where a user has lost control of their broadcast due to severe injury, medical emergency, police action, or being targeted with serious violence.
<b>Hate Speech and Acts of Aggression</b>	0.29%	0.22%	We do not tolerate conduct or speech that is hateful or that encourages or incites others to engage in hateful conduct. This includes inciting targeted community abuse, and expressions of hatred based on an identity-based protected characteristic.
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic, or repulsive content</b>	0.03%	0.06%	We don't permit streamers to be fully or partially nude. Additionally, content that exclusively focuses on extreme or gratuitous gore and violence is prohibited.





## Question 2: How safe is the platform for advertisers?

### Authorized Metric: Advertising Safety Error Rate

Twitch measures advertising safety error rate as a % of total advertising impressions delivered on content violative of the Twitch [Community Guidelines](#). We use the same methodology as that for violative view rate by aggregating impressions delivered on the day when a channel receives a violation.

GARM Category	Latest Period % of total Impressions	Previous Period % of total Impressions	Commentary
<b>Online Piracy</b>	<0.01%	<0.01%	We only allow sharing of content that streamers own, or otherwise have rights to or are authorized to share on Twitch. We do not allow pirated games or content from unauthorized private servers, movies, television shows, or sports matches, music streamers do not own the rights to share, goods or services protected by trademark, or other Twitch streamers' content if the steamer does not have authorization.
<b>Illegal Drugs / Tobacco / e-cigarettes / Vaping / Alcohol</b>	<0.01%	<0.01%	We do not permit any activity that may endanger a streamer's life or lead to physical harm. This includes illegal use of drugs and dangerous consumption of alcohol.
<b>Spam or Harmful Content</b>	0.45%	0.03%	We prohibit disruptive activities such as spamming, because these types of activities violate the integrity of Twitch services, and diminish users' experiences on Twitch.  We expect to see fluctuations in this category over time as we conduct ongoing audits and take action to remove bad actors.
<b>Terrorism</b>	<0.01%	<0.01%	We do not allow content that depicts, glorifies, encourages, or supports terrorism, or violent extremist actors or acts. This includes threatening to or encouraging others to commit acts that would result in serious physical harm to groups of people or significant property destruction. This metric includes the display or linking of terrorist or extremist propaganda, including graphic pictures or footage of terrorist or extremist violence, even for the purposes of denouncing such content.
<b>Other Violations</b>	0.08%	0.05%	For more information, on content that violates the Twitch guidelines, as well as more detailed takedown rates, please see our <a href="#">Community Guidelines</a> and <a href="#">Transparency Report</a> .



### Question 3: How effective is the platform in policy enforcement?

#### Authorized Metric: Total Enforcement Actions

Twitch measures our safety efforts as the total number of enforcements issued.

GARM Category	Latest Period Enforcement Actions	Previous Period Enforcement Actions	Commentary
<b>Adult &amp; Explicit Sexual Content</b>	39,118	35,359	<p>We prohibit content that involves nudity, and sexually explicit content. These are standards across live, image, and game content.</p> <p>The Law Enforcement Response (LER) team is an incredibly important part of the Trust &amp; Safety org. We continue to prioritize their work and invest in their training which has enabled us to better scale these types of investigations and identify more victims and offenders with each case, which promotes a safer service overall.</p>
<b>Crime and Harmful Acts to Individuals and Society, Human Right Violations</b>	83,298	80,406	<p>We prohibit content and conduct that may be upsetting or damaging, including content that focuses on violence, sexual violence, violent threats, self-harm behaviors, animal cruelty, dangerous or distracted driving, and other illegal, disturbing or frightening content/conduct. For a complete breakdown of our hate and harassment reports and enforcement, please see our <a href="#">Transparency Report</a>.</p>
<b>Death, Injury or Military Conflict</b>	8,237	4,006	<p>We prohibit content that is upsetting or damaging, including media and conduct that focuses on extreme gore, violence, or violent threats. We may temporarily remove the channel and associated content in situations where a user has lost control of their broadcast due to severe injury, medical emergency, police action, or being targeted with serious violence.</p> <p>Continued coverage of Ukraine and changes in our self-harm guidelines have contributed to a spike in enforcements related to this category.</p>
<b>Hate Speech and Acts of Aggression</b>	119,926	115,578	<p>We do not tolerate conduct or speech that is hateful or that encourages or incites others to engage in hateful conduct. This includes inciting targeted community abuse, and expressions of hatred based on an identity-based protected characteristic.</p>
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic, or repulsive content</b>	21,697	31,723	<p>We do not permit streamers to be fully or partially nude. Additionally, content that exclusively focuses on extreme or gratuitous gore and violence is prohibited.</p>



### Question 3: How effective is the platform in policy enforcement?

#### Authorized Metric: Total Enforcement Actions

Twitch measures our safety efforts as the total number of enforcements issued.

GARM Category	Latest Period Enforcement Actions	Previous Period Enforcement Actions	Commentary
Online Piracy	1,277	513	We only allow sharing of content that streamers own, or otherwise have rights to or are authorized to share on Twitch. We do not allow pirated games or content from unauthorized private servers, movies, television shows, or sports matches, music streamers do not own the rights to share, goods or services protected by trademark, or other Twitch streamers' content if the steamer does not have authorization.
Illegal Drugs / Tobacco / E-cigarettes / Vaping / Alcohol	41	17	We do not permit any activity that may endanger someone's life or lead to physical harm. This includes illegal use of drugs and dangerous consumption of alcohol.
Spam or Harmful Content	10,684,359	1,047,949	<p>We expect to see large fluctuations in this category over time depending on our cadence on taking mass action to remove large swathes of bad actors.</p> <p>Twitch provides tools such as customizable Blocked Terms and AutoMod, which allow channels to apply filters that proactively screen messages out of chat before they are seen. Channel moderators also actively monitor chat and can delete harmful or disruptive messages within seconds after they are posted.</p> <p>Additionally, Twitch programmatically identifies large bot accounts and takes bulk actions to enforce on not only bot accounts but also any associated account that might be participating in harmful behaviors. We expect to see fluctuations in this category over time, as we continue to take action against bot accounts.</p>
Terrorism	102	171	<p>We do not allow content that depicts, glorifies, encourages, or supports terrorism, or violent extremist actors or acts. This includes threatening to or encouraging others to commit acts that would result in serious physical harm to groups of people or significant property destruction. This metric includes the display or linking of terrorist or extremist propaganda, including graphic pictures or footage of terrorist or extremist violence, even for the purposes of denouncing such content.</p> <p>With the update to our Usernames Policy, usernames that glorify or promote acts of terrorism or terrorists now count as terrorism violations.</p>
Other Violations	341,866	502,317	For more information, on content that violates the Twitch guidelines, as well as more detailed takedown rates, please see our <a href="#">Community Guidelines</a> and <a href="#">Transparency Report</a> .



### Question 4: How does the platform perform at correcting mistakes

#### Authorized Metric: Total Enforcement Actions

The following metrics cover accounts that are acted upon and then appealed by users, and the decision to reinstate the account.

GARM Category	Latest Period		Previous Period	Commentary
	Appeal Rate	Reinstatement Rate	Appeal and Reinstatement Rate	
Adult & Explicit Sexual Content	4.82%	4.55%	6.13%    1.76%	<p>In H1 2022, we invested in an appeals portal for users and better internal tooling for our specialists to make filing and processing appeals easier for everyone.</p> <p>In H2 2022, our volume of appeals have decreased and we are steadily working through a backlog of appeals. We expect these rates to stabilize in the coming months.</p>
Arms & Ammunition				
Crime and Harmful Acts to Individuals and Society, Human Right Violations				
Death, Injury or Military Conflict				
Online piracy				
Hate speech and acts of aggression				
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust				
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol				
Spam or Harmful Content				
Terrorism				
Debated Sensitive Social Issue				





## Keeping LinkedIn Safe, Trusted, and Professional

At LinkedIn, we work to maintain a safe, trusted, and professional environment for all members and customers. LinkedIn is a professional network where members come together to find jobs, learn new skills, and build relationships. We connect a global community of professionals and companies to economic opportunity, within a safe, credible and transparent environment. To ensure conversations remain in the professional spirit of our platform, we have [Professional Community Policies](#) that apply to all content on our platform. In addition, our [Advertising Policies](#) provide advertisers with guidance on the types of ad content that is prohibited and restricted, and guidance on what to do if your ads are rejected.

Our Trust and Safety teams work diligently to keep content that violates these policies off of our platform. We have a [multidimensional approach](#) within our ecosystem to help us filter out violative content, which includes a combination of automatic and human-led detection.

We are proud to transparently share the efficacy of our efforts in enforcing these policies. Our semi-annual [Community Report](#) provides insights into the safety of the LinkedIn platform. In addition to content violations, this report also covers how we handle fake accounts, spam and scams, and copyrighted materials.

Our commitment to responsibility extends to our customers, and we have purpose-built solutions for our advertisers. For example, we monitor brand safety in the LinkedIn Feed via a score that measures the percentage of ads on LinkedIn Feed that appear next to detected violative content. **Since the creation of this metric in October 2021, the brand safety score has trended at or above 99% safe.**

## Additional Topics

### Jobs and Skills

Our value lies in our ability to connect professionals and employers around the world while identifying and reporting on fundamental economic trends such as in-demand jobs and skills, hiring patterns and the future of work. We are anchored in our vision to create economic opportunity for every member of the global workforce - and skills are the way to get there. We are dedicated to bringing skilling resources to everyone, everywhere and at every level, so that those unemployed can get back to work. We offer free access to training and tools that job seekers need, including placement and interview skills that help job seekers through the last mile of landing a job.

### Privacy

We understand how important the information people share on LinkedIn is to their professional lives. That is why we provide every member with clarity, consistency and control over the data they share and give all of our members the right to access, edit and delete their data at any time.

We enabled Group Identity, a solution that harnesses LinkedIn's first-party data to group members based on shared professional attributes - like seniority and industry - to help customers continue to target, measure, and optimize with efficiency and without the need for individual identifiers.

### Surfacing Safe and Constructive Conversations

In order to surface the safe, trusted, and professional content members expect on their LinkedIn Feed, our team uses [algorithms](#) responsibly. We're continuing to work on ways to improve the content members experience on LinkedIn, surfacing authentic, relevant, and substantive conversations to help members grow as professionals. To help drive [equitable outcomes](#) for all members, we continue to develop and share methods for assessing and [mitigating potential unfair bias](#) in our AI models.



### Methodology for Metrics

To find the latest on LinkedIn's transparency efforts visit [Our Transparency Center](#). It includes information on:

- Our [approach to keeping members safe](#)
- Our [approach to keeping brands safe](#)
- Our [approach to government requests](#)
- Our [Professional Community Policies](#)
- Our [Community Report](#)

### Brand Safety Score: How safe is advertising on LinkedIn Feed?

This metric shows the brand safety score of serving ads on LinkedIn Feed e.g., 99%+ safe means <1% of ads appeared next to detected policy violating content on LinkedIn Feed as defined by [LinkedIn Professional Community Policies](#). The score is derived using the following methodology:

- “100 minus {Number of ad impressions immediately above or below detected [violative content](#) in the Feed ([as determined by our Content Abuse Defense system](#)) divided by total number of Feed ad impressions.}”

For more details see [here](#).

### Removed Content: How much content did we remove?

- LinkedIn uses a combination of automatic prevention and human-led detection to remove content that violates our LinkedIn Professional Community Policies (see [here](#) for more information)
- Content Removed includes both public and private content removals on LinkedIn (such as posts, comments, and messages)

### Appealed Content: How much content did we remove that members appealed?

- The total number of member generated posts and comments (public content) that members appeal after we have removed the content for violating the LinkedIn Professional Community Policies
- All appeals are reviewed and the initial content removal decision is either upheld or reversed

### Reinstated Content: How much content did we reinstate after a member appeal?

- The total number of member generated posts and comments (public content) that LinkedIn restored after we originally removed the content for violating the LinkedIn Professional Community Policies
- We report only content that was restored in direct response to a member appeal



## Question 2: How safe is the platform for advertisers?

### Authorized Metric: Brand Safety Score

This metric shows the brand safety score of serving ads on LinkedIn Feed e.g., 99%+ safe means <1% of ads appeared next to detected policy violating content on LinkedIn Feed as defined by [LinkedIn Professional Community Policies](#).

**Comment:** We measure violative content by counting content that is reported by our members or flagged by our automated detection tools, and that is found to violate LinkedIn's Professional Community Policies. This methodology does not include violative content that is not reported or flagged, and may potentially be an undercount.

GARM Metric	Latest Period Q3 & Q4 2022	Previous Period Q1 & Q2 2022
	% brand safe impressions	% brand safe impressions
Brand Safety Score	99%+	99%+



### Question 3: How effective is the platform in enforcing its safety policy?

#### Authorized Metric: Content Removed

Violating content removed by LinkedIn. Each violative category includes both public and private content removals on LinkedIn.

#### [LinkedIn Professional Community Policies](#)

- Policies that govern content allowed on LinkedIn
- Enforcement of these policies is reflected in our biannual [Community Report](#)

LinkedIn Policy	Latest Period Q3 & Q4 2022	Previous Period Q1 & Q2 2022
	Content Removed	Content Removed
<a href="#">Nudity and Adult</a>	46,401	34,163
Child Exploitation	274	1,663
<a href="#">Harassment or abusive</a>	204,365	178,926
<a href="#">Hateful or derogatory</a>	56,821	37,835
<a href="#">Violent or graphic</a>	61,333	60,990
<a href="#">Copyright policies</a>	1,098	1,012
Other ( <a href="#">misinformation</a> )	137,988	172,387
<a href="#">Nudity and Adult</a>	46,401	34,163





## Question 4: How responsive is the platform at correcting mistakes?

**Authorized Metric:** Appeals, Reinstatements

Content appealed after human reviewed removal action; Content reinstated after a member appeal

**Comment:** Our metric only counts appeals resulting from human reviewed content.

GARM Metric	Latest Period Q3 & Q4 2022	Previous Period Q1 & Q2 2022
Total Content Appeals	23,625	30,656
Total Content Reinstatements	4,273	5,087



# Mapping GARM Categories to LinkedIn Professional Community Policies

In the table below, we have mapped each of the GARM Brand Safety Floor Categories to the closest corresponding LinkedIn Professional Community Policy(s). We offer this table to help you understand how our Professional Community Policies compare with GARM's definitions of brand unsafe content. These policies apply to all member and customer generated content on LinkedIn, including articles, messages, images, videos and ads.

GARM Brand Safety Floor Category + Definition	Relevant <a href="#">LinkedIn Professional Community Policy</a>
<b>Adult &amp; Explicit Sexual Content</b> <ul style="list-style-type: none"> <li>• Illegal sale, distribution, and consumption of child pornography</li> <li>• Explicit or gratuitous depiction of sexual acts, and/or display of genitals, real or animated</li> </ul>	<p><b><a href="#">Nudity and Adult</a></b> Do not share material depicting nudity or sexual activity.</p> <p><b><a href="#">Child Exploitation</a></b> Do not share material depicting the exploitation of children: We have zero tolerance for content that depicts the sexual exploitation of children. Do not share, post, transmit, or solicit child exploitation material through or using our platform.</p>
<b>Arms &amp; Ammunition</b> <ul style="list-style-type: none"> <li>• Promotion and advocacy of Sales of illegal arms, rifles, and handguns</li> <li>• Instructive content on how to obtain, make, distribute, or use illegal arms</li> <li>• Glamorization of illegal arms for the purpose of harm to others</li> <li>• Use of illegal arms in unregulated environments</li> </ul>	<p><b><a href="#">Illegal, dangerous, and inappropriate commercial activity</a></b> Do not promote, sell or attempt to purchase illegal or dangerous goods or services. We don't allow content that facilitates the purchase of illegal or dangerous goods and/or services, prostitution, and escort services.</p>
<b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b> <ul style="list-style-type: none"> <li>• Graphic promotion, advocacy, and depiction of willful harm and actual unlawful criminal activity – Explicit violations/demeaning offenses of Human Rights (e.g., human trafficking, slavery, self-harm, animal cruelty etc.),</li> <li>• Harassment or bullying of individuals and groups</li> </ul>	<p><b><a href="#">Violent or Graphic</a></b> Do not threaten, incite, or promote violence: We don't allow threatening or inciting violence of any kind. We don't allow individuals or groups that engage in or promote violence, property damage, or organized criminal activity.</p> <p><b><a href="#">Illegal, dangerous, and inappropriate commercial activity</a></b> Do not promote, sell or attempt to purchase illegal or dangerous goods or services. We don't allow content that facilitates the purchase of illegal or dangerous goods and/or services, prostitution, and escort services.</p> <p><b><a href="#">Harassment and Abusive Content</a></b> Do not post harassing content: We don't allow bullying or harassment. This includes targeted personal attacks, intimidation, shaming, disparagement, and abusive language directed at other members.</p> <p><b><a href="#">Sexual Innuendo and Unwanted Advances</a></b> Do not engage in sexual innuendos or unwanted advances. We don't allow unwanted expressions of attraction, desire, requests for romantic relationships, marriage proposals, sexual advances or innuendo, or lewd remarks.</p> <p><b><a href="#">Scams and Fraud</a></b> Do not scam, defraud, deceive others. Do not use LinkedIn to facilitate romance scams, promote pyramid schemes, or otherwise defraud members. Do not share malicious software that puts our members, platform, or services at risk.</p>
<b>Death, Injury or Military Conflict</b> <ul style="list-style-type: none"> <li>• Promotion, incitement or advocacy of violence, death or injury</li> <li>• Murder or Willful bodily harm to others</li> <li>• Graphic depictions of willful harm to others</li> <li>• Incendiary content provoking, enticing, or evoking military aggression</li> <li>• Live action footage/photos of military actions &amp; genocide or other war crimes</li> </ul>	<p><b><a href="#">Violent or Graphic</a></b> Do not threaten, incite, or promote violence: We don't allow threatening or inciting violence of any kind. We don't allow individuals or groups that engage in or promote violence, property damage, or organized criminal activity.</p>



# Mapping GARM Categories to LinkedIn Professional Community Policies

GARM Brand Safety Floor Category + Definition	Relevant <a href="#">LinkedIn Professional Community Policy</a>
<b>Online piracy</b> <ul style="list-style-type: none"> <li>Pirating, Copyright infringement, &amp; Counterfeiting</li> </ul>	<p><a href="#">Copyrighted Materials</a></p> <p><a href="#">Illegal, dangerous, and inappropriate commercial activity</a> Do not promote, sell or attempt to purchase illegal or dangerous goods or services. We don't allow content that facilitates the purchase of illegal or dangerous goods and/or services, prostitution, and escort services.</p>
<b>Hate speech &amp; acts of aggression</b> <ul style="list-style-type: none"> <li>Behavior or content that incites hatred, promotes violence, vilifies, or dehumanizes groups or individuals based on race, ethnicity, gender, sexual orientation, gender identity, age, ability, nationality, religion, caste, victims and survivors of violent acts and their kin, immigration status, or serious disease sufferers.</li> </ul>	<p><a href="#">Hateful or Derogatory</a> Do not be hateful. We don't allow content that attacks, denigrates, intimidates, dehumanizes, incites or threatens hatred, violence, prejudicial or discriminatory action against individuals or groups because of their actual or perceived race, ethnicity, national origin, caste, gender, gender identity, sexual orientation, religious affiliation, age, or disability status. Hate groups are not permitted on LinkedIn.</p> <p><a href="#">Harassment and Abusive Content</a> Do not post harassing content: We don't allow bullying or harassment. This includes targeted personal attacks, intimidation, shaming, disparagement, and abusive language directed at other members.</p>
<b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b> <ul style="list-style-type: none"> <li>Excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult.</li> </ul>	<p><a href="#">Hateful or Derogatory</a> Do not be hateful. We don't allow content that attacks, denigrates, intimidates, dehumanizes, incites or threatens hatred, violence, prejudicial or discriminatory action against individuals or groups because of their actual or perceived race, ethnicity, national origin, caste, gender, gender identity, sexual orientation, religious affiliation, age, or disability status. Hate groups are not permitted on LinkedIn.</p> <p><a href="#">Violent or Graphic</a> Do not threaten, incite, or promote violence: We don't allow threatening or inciting violence of any kind. We don't allow individuals or groups that engage in or promote violence, property damage, or organized criminal activity.</p>
<b>Illegal Drugs/Tobacco/ecigarettes/Vaping/Alcohol</b> <ul style="list-style-type: none"> <li>Promotion or sale of illegal drug use – including abuse of prescription drugs. Federal jurisdiction applies, but allowable where legal local jurisdiction can be effectively managed</li> <li>Promotion and advocacy of Tobacco and e-cigarette (Vaping) &amp; Alcohol use to minors</li> </ul>	<p><a href="#">Illegal, dangerous, and inappropriate commercial activity</a> Do not promote, sell or attempt to purchase illegal or dangerous goods or services. We don't allow content that facilitates the purchase of illegal or dangerous goods and/or services, prostitution, and escort services.</p>
<b>Spam or Harmful Content</b> <ul style="list-style-type: none"> <li>Malware/Phishing</li> </ul>	<p><a href="#">Spam Content</a> Do not spam members or the platform. We don't allow untargeted, irrelevant, obviously unwanted, unauthorized, in appropriately commercial or promotional, or gratuitously repetitive messages or similar content.</p> <p><a href="#">Scams and Fraud</a> Do not scam, defraud, deceive others. Do not use LinkedIn to facilitate romance scams, promote pyramid schemes, or otherwise defraud members. Do not share malicious software that puts our members, platform, or services at risk.</p> <p><a href="#">Illegal, dangerous, and inappropriate commercial activity</a> Do not promote, sell or attempt to purchase illegal or dangerous goods or services. We don't allow content that facilitates the purchase of illegal or dangerous goods and/or services, prostitution, and escort services.</p>



# Mapping GARM Categories to LinkedIn Professional Community Policies

GARM Brand Safety Floor Category + Definition	Relevant <a href="#">LinkedIn Professional Community Policy</a>
<p><b>Terrorism</b></p> <ul style="list-style-type: none"> <li>Promotion and advocacy of graphic terrorist activity involving defamation, physical and/or emotional harm of individuals, communities, and society</li> </ul>	<p><a href="#">Dangerous organizations and individuals</a></p> <p>Do not share content promoting dangerous organizations or individuals. We don't allow any terrorist organizations or violent extremist groups on our platform. And we don't allow any individuals who affiliate with such organizations or groups to have a LinkedIn profile. Content that depicts terrorist activity, that is intended to recruit for terrorist organizations, or that threatens, promotes, or supports terrorism in any manner is not tolerated.</p>
<p><b>Debated Sensitive Social Issue</b></p> <ul style="list-style-type: none"> <li>Insensitive, irresponsible and harmful treatment of debated social issues and related acts that demean a particular group or incite greater conflict;</li> </ul>	<p><a href="#">Hateful or Derogatory</a></p> <p>Do not be hateful. We don't allow content that attacks, denigrates, intimidates, dehumanizes, incites or threatens hatred, violence, prejudicial or discriminatory action against individuals or groups because of their actual or perceived race, ethnicity, national origin, caste, gender, gender identity, sexual orientation, religious affiliation, age, or disability status. Hate groups are not permitted on LinkedIn.</p>
	<p><a href="#">Harassment and Abusive</a></p> <p>Do not post harassing content: We don't allow bullying or harassment. This includes targeted personal attacks, intimidation, shaming, disparagement, and abusive language directed at other members.</p>
<p><b>Other</b></p>	<p><a href="#">Misinformation</a></p> <p>Do not share false or misleading content: Do not share content in a way that you know is, or think may be, misleading or inaccurate, including misinformation or disinformation.</p>

## Appendices & FAQ

### How is the report created and what is the governance?

As this is an aggregated report, the metrics and measures are sourced from existing first-party transparency reports that are already produced by the GARM platforms that have opted to participate in the report. The Aggregated Report is an abridged version of those as it streamlines the current reporting practices into a framework that is relevant and useful to advertisers.

#### STEP 1:

Platforms involved in GARM confirm participation

#### STEP 2:

GARM Working Group distributes data submission and commentary submission template

#### STEP 3:

WFA aggregates submissions and GARM Steer Team develops analysis for Executive Summary

#### STEP 4:

GARM platforms review and confirm content for accuracy and GARM Working Group approves content

#### STEP 5:

WFA GARM publishes report

The GARM Steer Team and GARM Initiative Lead are accountable for the final decisions on the report, corresponding to overall GARM Governance, detailed on the GARM section of the WFA website.

### Why are we focusing on these four core questions?

After a thorough review and discussion, the GARM Measurement & Oversight Working determined there are three perspectives to take into account when measuring harmful content: consumer experience, advertiser experience, and platform actions.

From there we were able to identify the questions that best help us assess the size of the challenge and that the best approach to structuring a measurement solution would be based on a series of questions that would size the challenge in a consumer-centric and advertiser-centric way and show platform progress against it.

PERSPECTIVE	AREA FOR ANALYSIS	CORE QUESTION
Consumer experience	Amount of harmful content getting thru to consumers	How safe is the platform for consumers?
Advertiser experience	Amount of advertising inadvertently placed next to harmful content	How safe is the platform for advertisers?
Platform actions and progress	Ability of the platform to take action on harmful content and how many times it has been viewed by consumers Ability of the platform to manage the need for an open and safe communications experience	How effective is the platform in enforcing its safety policies? How responsive is the platform in correcting mistakes?

## Appendices & FAQ

These four core questions were reviewed by the GARM Steer Team and the GARM Community and endorsed as the means to structure the report and identify appropriate measures.

### What are ‘Authorized Metrics’ and how were they identified?










Authorized Metrics are a set of measures that the GARM Measurement & Oversight Working Group identified in their review of current measurement techniques. The Working Group reviewed a series of 80 candidate measures for the four core questions. In discussions, the group concluded that certain measures could represent a more suitable way to answer the question while advancing methodological best practices. The candidate measures for authorized metrics were reviewed by the GARM Steer Team and along with the MRC (Media Ratings Council).

The following table details the authorized metrics per question for the GARM Aggregated Measurement Report:










CORE QUESTION	AUTHORIZED METRIC	DEFINITION + OVERVIEW	RATIONALE
How safe is the platform for consumers?	Prevalence of violating content or Violative View Rate	The percentage of views that contain content that is deemed as violative	Establishes a ratio based on typical user content consumption. Prevalence or Violative View Rate examines views of unsafe/violating content as a proportion of all views.
How safe is the platform for advertisers?	Prevalence of violating content or Advertising Safety Error Rate	The percentage of views that contain content that is deemed as violative  The percentage of views of monetized content that contain violative content	Monetization prevalence examines unsafe content viewed as a proportion of monetized content viewed
How effective is the platform in policy enforcement?	Removals of Violating Content + Removal of Violating Accounts Removals of Violating Content expressed by how many times it has been viewed	Pieces of violating content removed  Accounts removed due to repeat policy violation  Pieces of violating content removed categorized by how many times they were viewed by users	Platform teams spend a considerable amount of time removing violating content and bad actors from their platforms – the magnitude of the efforts should be reported to marketers. It is also important to marketers to understand how many times harmful content has been removed.
How does the platform perform at correcting mistakes?	Appeals Reinstatements	Number of pieces of violating content removed that are appealed  Number of pieces of violating content removed that are appealed and then reinstated	Platform should be responsive to their users and policy should be consistent with a policy of free and safe speech. For this reason we look at appeals and reinstatement of content removed.

In the event a platform is unable to submit a question response with an authorized metric, they are encouraged to submit a next best measure. Inclusion does not represent GARM endorsement of the measure, but it allows platforms to present how they currently answer the GARM Aggregated Measurement Report’s questions in the ways which they have developed individually.

**The next table provides an overview of platform submission of data for Volume 2:**

										
How safe is the platform for consumers?	Prevalence Violative View Rate	Authorized Metric	Authorized Metric	Authorized Metric	Next Best Measure	Next Best Measure	Next Best Measure	Authorized Metrics	Next Best Measure	Not Submitted
How safe is the platform for advertisers?	Advertiser Safety Error Rate or Prevalence	Authorized Metric	Authorized Metric	Authorized Metric	Next Best Measure	Next Best Measure	Next Best Measure	Authorized Metric	Authorized Metric	Authorized Metric
How effective is the platform at enforcing its safety policies?	Removals of violating content	Authorized Metric	Authorized Metric	Authorized Metric	Authorized Metric	Next Best Measure	Authorized Metric	Authorized Metric	Authorized Metric	Authorized Metric
	Removal of violating accounts by views	Authorized Metric	Authorized Metric	Not Submitted	Next Best Measure	Authorized Metric	Authorized Metric	Authorized Metric	Authorized Metric	Not Submitted
	Removal of violating accounts	Authorized Metric	Authorized Metric	Not Submitted	Authorized Metric	Not Submitted	Authorized Metric	Authorized Metric	Authorized Metric	Not Submitted
How responsive is the platform in correcting mistakes?	Appeals (pieces of content)	Authorized Metric	Authorized Metric	Authorized Metric	Not Submitted	Not Submitted	Authorized Metric	Not Submitted	Authorized Metric	Authorized Metric
	Reinstatements (pieces of content)	Authorized Metric	Authorized Metric	Authorized Metric	Not Submitted	Not Submitted	Authorized Metric	Not Submitted	Authorized Metric	Authorized Metric

## Aggregated Measurement Report Volume 4: Date ranges for platform data submitted

	Q3 2021	Q4 2021	Q1 2022	Q2 2022	Q3 2022	Q4 2022
			PREVIOUS PERIOD		LATEST PERIOD	
			PREVIOUS PERIOD		LATEST PERIOD	
			PREVIOUS PERIOD		LATEST PERIOD	
		PREVIOUS PERIOD		LATEST PERIOD		
			PREVIOUS PERIOD		LATEST PERIOD	
			PREVIOUS PERIOD		LATEST PERIOD	
		PREVIOUS PERIOD		LATEST PERIOD		
			PREVIOUS PERIOD		LATEST PERIOD	
			PREVIOUS PERIOD		LATEST PERIOD	



### Is the data featured in the GARM Aggregated Measurement Report audited?

No; the source data for the reports is not audited at this stage. The Aggregated Measurement Report is built from platform first-party transparency report data. Within GARM there is an understood goal to have these reports audited by independent parties, such as the MRC and other auditing firms. This process is ongoing, and we recognize efforts underway with specific platforms. The progress of auditing the first-party transparency reporting is being tracked and assessed by the GARM Steer Team, the MRC, and the individual platforms. The GARM Steer Team and its sponsors have communicated the need to audit activities across brand safety controls, brand safety measurement, brand safety integrations and first-party transparency reporting. GARM reports on the progress of these audits to its members and its executive stakeholders.

### There are currently three levels of audits being pursued within GARM that have been prioritized by the GARM Steer Team:

**Level 1:** Brand Safety Controls & Measurement

**Level 2:** Brand Safety Integrations

**Level 3:** Brand Safety Transparency Reporting

Each GARM platform is managing their respective agreement and roadmap for audits and communicating progress to the GARM Steer Team. An update of this process will be in upcoming GARM Quarterly Updates. It is important to note that currently no platform has an externally audited Transparency Report.

### How often does the report come out and how is it created?

The GARM Aggregated Measurement Report is issued twice a year, using each participating platform's first-party reporting data, and references two time periods – latest 6 months, and prior 6 months as a trended reference period. Where platforms currently report quarterly, each quarter is reported separately within these two time periods.

The report is created within GARM and uses first-party reporting data sources as its basis. The data relevant to the core questions are collected by GARM in a template issued to reporting platforms that allow for both the reporting of metrics and explanation of measures and changes. The templates are then consolidated into a chapter. GARM then provides commentary on industry improvement opportunities, highlights steps that are successful, and acknowledges best-in-class steps by individual players.

The GARM Aggregated Measurement Report is created by using established first party safety and transparency reports, which are reflective of individual platform policies and their enforcement. The metrics presented indicate the presence of content that violates platform policies and actions taken by the platforms against the violating content. The comparative framework uses GARM categories for the monetization of harmful content, Platform policies were mapped to this GARM categorization and then agreed. An overview of the results of this process can be found below:

GARM Aggregated Measurement Report

GARM Content Category	Relevant Platform Policy								
	YouTube	Facebook	Instagram	Twitter	TikTok	Pinterest	Snap	Twitch	LinkedIn
Adult & Explicit Sexual Content	<ul style="list-style-type: none"> <li>Nudity &amp; Sexual Content</li> <li>Child Safety</li> </ul>	<ul style="list-style-type: none"> <li>Adult Nudity and Sexual Activity,</li> <li>Child Sexual Exploitation, Abuse and Nudity,</li> <li>Sexual Solicitation</li> </ul>	<ul style="list-style-type: none"> <li>Adult Nudity and Sexual Activity,</li> <li>Child Sexual Exploitation, Abuse and Nudity,</li> <li>Sexual Solicitation</li> </ul>	<ul style="list-style-type: none"> <li>Non-Consensual Nudity</li> <li>Sensitive Media</li> <li>Child Sexual Exploitation</li> </ul>	<ul style="list-style-type: none"> <li>Minor safety – sexual exploitation of minors</li> <li>Adult nudity and sexual activities</li> </ul>	<ul style="list-style-type: none"> <li>Adult Sexual Services</li> <li>Adult Content</li> </ul>	<ul style="list-style-type: none"> <li>Sexually Explicit Content</li> </ul>	<ul style="list-style-type: none"> <li>Nudity, Pornography, and Other Sexual Content</li> </ul>	<ul style="list-style-type: none"> <li>Nudity and Adult</li> <li>Child Exploitation</li> </ul>
Arms & Ammunition	<ul style="list-style-type: none"> <li>Firearms</li> </ul>	<ul style="list-style-type: none"> <li>Violence and Incitement</li> <li>Restricted Goods and Services</li> </ul>	<ul style="list-style-type: none"> <li>Violence and Incitement</li> <li>Restricted Goods and Services</li> </ul>	<ul style="list-style-type: none"> <li>Illegal or certain regulated good or services</li> </ul>	<ul style="list-style-type: none"> <li>Illegal activities and regulated goods – weapons</li> </ul>	<ul style="list-style-type: none"> <li>Dangerous Goods and Activities</li> </ul>	<ul style="list-style-type: none"> <li>Regulated Goods</li> </ul>	<ul style="list-style-type: none"> <li>Violence and Threats</li> </ul>	<ul style="list-style-type: none"> <li>Illegal, dangerous, and inappropriate commercial activity</li> </ul>
Crime & Harmful acts to individuals and Society, Human Right Violations	<ul style="list-style-type: none"> <li>Harmful or Dangerous Content</li> <li>Hate Speech</li> <li>Harassment or cyberbullying</li> </ul>	<ul style="list-style-type: none"> <li>Adult Nudity and Sexual Activity</li> <li>Violence and Incitement</li> <li>Bullying and Harassment</li> <li>Violent and Graphic Content</li> <li>Child Sexual Exploitation, Abuse and Nudity</li> <li>Suicide and Self-Injury</li> <li>Dangerous Individuals and Organizations</li> <li>Restricted Goods and Services</li> </ul>	<ul style="list-style-type: none"> <li>Adult Nudity and Sexual Activity</li> <li>Violence and Incitement</li> <li>Bullying and Harassment</li> <li>Violent and Graphic Content</li> <li>Child Sexual Exploitation, Abuse and Nudity</li> <li>Suicide and Self-Injury</li> <li>Dangerous Individuals and Organizations</li> <li>Restricted Goods and Services</li> </ul>	<ul style="list-style-type: none"> <li>Violence</li> <li>Abuse and harassment</li> </ul>	<ul style="list-style-type: none"> <li>Illegal activities and regulated goods -criminal activities</li> </ul>	<ul style="list-style-type: none"> <li>Child Sexual Exploitation</li> <li>Self-Harm</li> <li>Harassment &amp; Criticism</li> </ul>	<ul style="list-style-type: none"> <li>Threatening / Violence / Harm:</li> </ul>	<ul style="list-style-type: none"> <li>Self-Destructive Behaviour</li> <li>Hateful Conduct and Harassment</li> </ul>	<ul style="list-style-type: none"> <li>Violent or graphic</li> <li>Illegal, dangerous and inappropriate commercial activity</li> <li>Harassment and Abusive Content</li> <li>Sexual Innuendo and Unwanted Advances</li> <li>Scams and fraud</li> </ul>
Death, Injury or Military Conflict	<ul style="list-style-type: none"> <li>Violent or Graphic Content</li> <li>Harmful or Dangerous Content</li> <li>Suicide &amp; Self-Injury</li> </ul>	<ul style="list-style-type: none"> <li>Violence and Incitement</li> <li>Violent and Graphic Content</li> <li>Suicide and Self-Injury</li> </ul>	<ul style="list-style-type: none"> <li>Violence and Incitement</li> <li>Violent and Graphic Content</li> <li>Suicide and Self-Injury</li> </ul>	<ul style="list-style-type: none"> <li>Promoting Self-harm</li> </ul>	<ul style="list-style-type: none"> <li>Violent and Graphic Content</li> </ul>	<ul style="list-style-type: none"> <li>Graphic Violence and Threats</li> </ul>	<ul style="list-style-type: none"> <li>Threatening / Violence / Harm</li> </ul>	<ul style="list-style-type: none"> <li>Violence and Threats</li> <li>Extreme Violence, Gore, and Other Obscene Content</li> </ul>	<ul style="list-style-type: none"> <li>Violent or graphic</li> </ul>
Online piracy	<ul style="list-style-type: none"> <li>Fake Engagement</li> <li>Impersonation</li> <li>Sale of illegal or regulated goods or services</li> <li>YouTube Terms of Service</li> </ul>	<ul style="list-style-type: none"> <li>Intellectual Property</li> <li>Copyright</li> <li>Intellectual Property Counterfeit</li> <li>Intellectual Property Trademark</li> </ul>	<ul style="list-style-type: none"> <li>Intellectual Property</li> <li>Copyright</li> <li>Intellectual Property Counterfeit</li> <li>Intellectual Property Trademark</li> </ul>	<ul style="list-style-type: none"> <li>Copyright</li> <li>Trademark</li> </ul>	<ul style="list-style-type: none"> <li>Integrity and authenticity – intellectual property violations</li> </ul>	<ul style="list-style-type: none"> <li>Copyright</li> <li>Trademark</li> </ul>	<ul style="list-style-type: none"> <li>Spam</li> </ul>	<ul style="list-style-type: none"> <li>Spam, Scams, and Other Malicious Content</li> </ul>	<ul style="list-style-type: none"> <li>Copyrighted Materials</li> <li>Illegal, dangerous, and inappropriate commercial activity</li> </ul>
Hate speech & acts of aggression	<ul style="list-style-type: none"> <li>Hate Speech</li> </ul>	<ul style="list-style-type: none"> <li>Hate speech</li> <li>Bullying and Harassment</li> <li>Dangerous Individuals and Organizations</li> </ul>	<ul style="list-style-type: none"> <li>Hate speech</li> <li>Bullying and Harassment</li> <li>Dangerous Individuals and Organizations</li> </ul>	<ul style="list-style-type: none"> <li>Hateful Conduct</li> </ul>	<ul style="list-style-type: none"> <li>Hate Speech</li> <li>Hateful Behavior</li> </ul>	<ul style="list-style-type: none"> <li>Hateful Activities</li> </ul>	<ul style="list-style-type: none"> <li>Threatening / Violence / Harm</li> </ul>	<ul style="list-style-type: none"> <li>Hateful Conduct and Harassment</li> </ul>	<ul style="list-style-type: none"> <li>Hateful or Derogatory</li> <li>Harassment and Abusive Content</li> </ul>
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	<ul style="list-style-type: none"> <li>Violent or Graphic Content</li> <li>Age Restriction</li> </ul>	<ul style="list-style-type: none"> <li>Hate Speech</li> <li>Bullying and Harassment</li> </ul>	<ul style="list-style-type: none"> <li>Hate Speech</li> <li>Bullying and Harassment</li> </ul>	<ul style="list-style-type: none"> <li>Sensitive Media</li> </ul>	<ul style="list-style-type: none"> <li>Hateful Behavior – Slurs</li> <li>Harassment &amp; Bullying</li> </ul>	<ul style="list-style-type: none"> <li>Harassment &amp; Criticism</li> </ul>		<ul style="list-style-type: none"> <li>Extreme Violence, Gore, and Other Obscene Content</li> </ul>	<ul style="list-style-type: none"> <li>Hateful or Derogatory</li> <li>Violent or Graphic</li> </ul>
Illegal drugs, tobacco, e-cigarettes, vaping	<ul style="list-style-type: none"> <li>Sale of Illegal or Regulated Goods or Services</li> <li>Harmful or dangerous content</li> </ul>	<ul style="list-style-type: none"> <li>Regulated Goods: Drugs</li> </ul>	<ul style="list-style-type: none"> <li>Regulated Goods: Drugs</li> </ul>	<ul style="list-style-type: none"> <li>Illegal or certain regulated goods or services</li> </ul>	<ul style="list-style-type: none"> <li>Illegal activities and regulated goods – drugs, controlled substances, alcohol and tobacco</li> </ul>	<ul style="list-style-type: none"> <li>Dangerous Goods and Activities</li> </ul>	<ul style="list-style-type: none"> <li>Regulated Goods</li> </ul>	<ul style="list-style-type: none"> <li>Self-destructive behaviour</li> </ul>	<ul style="list-style-type: none"> <li>Illegal, dangerous, and inappropriate commercial activity</li> </ul>
Spam & Malware	<ul style="list-style-type: none"> <li>Spam, Deceptive Practices, scams, and misinformation</li> </ul>	<ul style="list-style-type: none"> <li>Spam</li> </ul>	<ul style="list-style-type: none"> <li>Spam</li> </ul>	<ul style="list-style-type: none"> <li>Private information</li> <li>Impersonation</li> <li>Platform manipulation</li> </ul>	<ul style="list-style-type: none"> <li>Integrity and authenticity – spam and fake engagement</li> </ul>	<ul style="list-style-type: none"> <li>Spam</li> </ul>	<ul style="list-style-type: none"> <li>Spam</li> </ul>	<ul style="list-style-type: none"> <li>Spam, Scams, and Other Malicious Content</li> </ul>	<ul style="list-style-type: none"> <li>Spam content</li> <li>Scams and Fraud</li> <li>Illegal, dangerous and inappropriate commercial activity</li> </ul>
Terrorism	<ul style="list-style-type: none"> <li>Violent criminal organizations</li> </ul>	<ul style="list-style-type: none"> <li>Dangerous Organizations: Terrorism</li> <li>Dangerous Organizations: Organized Hate</li> </ul>	<ul style="list-style-type: none"> <li>Dangerous Organizations: Terrorism</li> <li>Dangerous Organizations: Organized Hate</li> </ul>	<ul style="list-style-type: none"> <li>Terrorism or Violent Extremism</li> </ul>	<ul style="list-style-type: none"> <li>Violent Extremism</li> <li>Dangerous individuals and organizations – Terrorists and terrorist organizations</li> </ul>	<ul style="list-style-type: none"> <li>Violent Actors</li> </ul>	<ul style="list-style-type: none"> <li>Terrorism</li> </ul>	<ul style="list-style-type: none"> <li>Violence and Threats</li> </ul>	<ul style="list-style-type: none"> <li>Dangerous organizations and individuals</li> </ul>
Debated Sensitive Social Issues		<ul style="list-style-type: none"> <li>Hate Speech</li> <li>Bullying and Harassment</li> </ul>	<ul style="list-style-type: none"> <li>Hate Speech</li> <li>Bullying and Harassment</li> </ul>		<ul style="list-style-type: none"> <li>Hateful Behavior</li> </ul>	<ul style="list-style-type: none"> <li>Civic Misinformation</li> <li>Conspiracy Theories</li> <li>Medical Misinformation</li> <li>Climate Misinformation</li> </ul>			<ul style="list-style-type: none"> <li>Hateful or Derogatory</li> <li>Harassment or Abusive</li> </ul>
Other	<ul style="list-style-type: none"> <li>Any categories not specifically accounted for in the above (e.g. multiple policy violations)</li> </ul>	<ul style="list-style-type: none"> <li>COVID-19 and Vaccine Policy and Protections</li> </ul>	<ul style="list-style-type: none"> <li>COVID-19 and Vaccine Policy and Protections</li> </ul>	<ul style="list-style-type: none"> <li>Covid Integrity</li> <li>Covid-19 Misleading Information</li> </ul>				<ul style="list-style-type: none"> <li>Suspension Evasion</li> <li>Unauthorized Sharing of Private Information</li> <li>Impersonation</li> <li>Cheating in Online Games</li> </ul>	<ul style="list-style-type: none"> <li>Misinformation</li> </ul>



**World Federation of Advertisers**

London, Brussels, Singapore, New York

[wfanet.org](http://wfanet.org)

[info@wfanet.org](mailto:info@wfanet.org)

+32 2 502 57 40

twitter [@wfamarketers](https://twitter.com/wfamarketers)

[youtube.com/wfamarketers](https://youtube.com/wfamarketers)

[linkedin.com/company/wfa](https://linkedin.com/company/wfa)