GARM Global Alliance for Responsible Media

GARM Aggregated Measurement Report

Volume 4 | November 2022

Contents

GARM Global Alliance for Responsible Media

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.

GARM Global Alliance for Responsible Media

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 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 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 Oversights 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 and Twitch. For a more detailed overview of how we've worked within GARM to create this report, please see the Appendix.

Executive Summary

Now in our fourth volume of the GARM Aggregated Measurement Report, we are starting to see how changes in safety and enforcement data are driven by policy development, technology advances and real-world events. Volume four of the Aggregated Measurement Report reports on and analyzes participating platform data from Q3 2021 through Q4 2022. This new installment features the following key advancements for the report and underlying reporting efforts:

01

First, we have expanded GARM's Aggregated Measurement Reporting templates to include Misinformation. In June 2022, we updated the GARM Brand Safety Floor + Suitability Framework to include a definition for Misinformation. While this is a new definition, we expect to see a more structured and distinct reporting process by our members for this content category, and specific enforcement areas therein. As of this volume, all GARM members highlight ongoing enforcement of their work in this space in their platform commentaries. GARM members formally reporting on this content category at present volume include: YouTube, Twitter, Pinterest. We also anticipate that co-regulatory work via mechanisms such as the European Commission Code of Practice on Disinformation will also help develop enforcement and reporting in this area.

02

Second, we've continued to update the key information section of the document towards the Appendix that features key information relative to the time span of the data analyzed, map of how platform policies support the GARM Categories, and an overview of data shared by participating platforms.

In reviewing the data there are three learnings in this period

LEARNING 1:

GARM Categories Spam & Malware and Adult & Explicit Sexual Content continue to dominate enforcement in terms of volume. Adult & Explicit Sexual Content continues to be a high enforcement area, whether it is removal of content or accounts. Looking at Content Removals and Account Removals, Adult & Explicit accounts for the top removal reason for three of the submitting platforms (YouTube, Pinterest, Snap). We see similar volumes of Adult Content in account removals with Twitter, where Adult Content represents 45% of total account removals. When looking at Spam & Malware, that content category is the top removal reason for an additional two of the submitting platforms (Facebook, Twitter).
Proactive removal rates in these areas continue to be high, given the advances of technology-led blocking rates.

LEARNING 2:

Death, Injury, Military Conflict and Arms & Ammunition show highest category enforcement growth, likely due to the coverage of Russia's invasion of Ukraine. While our last volume in April was released after Russia's invasion of Ukraine, this is the first we are seeing the data feed into the report. Looking at Arms & Ammunition first, we see that enforcement for content removals in this category has increased in the latest period across 5-of-7 submitting platforms. Shifting to Death, Injury and Military Conflict, we see that this enforcement area has significantly increased for three platforms: Pinterest (+798%), Twitter (+23%), Facebook (+21%).

LEARNING 3:

Enforcement on Illegal Drugs, Tobacco, e-cigarettes, Vaping, Alcohol are significantly increasing. In half of the platforms participating in the Aggregated Measurement Report, we have seen an increase in content removals contrasting latest period versus prior period. Twitter reported the highest increase (+36%), followed by Instagram (+23%). Shifting to account removals, we see corresponding increases in enforcement across Twitter (+37%) and Pinterest (+37%). When consulted, some platforms noted refinements in enforcement guidelines and improvements in AI as a cause for the increased enforcement.

A look forward to Volume 5 and beyond:

Accreditation of Monetization Metrics and Transparency Reporting:

We continue to support all platform efforts to pursue independent accreditation of transparency reporting. We recognize that media buying customers are one stakeholder group and recognize the complementary regulatory ambitions and civil society organization requests here too.

For the purposes of monetization and transparency reporting, we still support MRC's update to the Content-Level Brand Safety Controls Audit specification which has monetization transparency as part of that scope of work. We will acknowledge independent audits to reported data in the Aggregated Measurement Report as they are accredited by MRC or other aligned auditing standard bodies.

YouTube thus far is the only platform to receive the full platform accreditation, inclusive of their monetization safety metric. Most recently Facebook has received MRC brand safety accreditation for a portion of their advertising products, and we look forward to their continued progress to reach full platform accreditation. Twitter is anticipated to proceed with their MRC brand safety accreditation starting next year.

 We call on all of our platforms to proceed with the MRC audit, given its ability to give needed assurance on advertising sales and safety practices. We eagerly await more platforms taking this step.

Increased Analysis:

Our next volume will have two years of data. This will give us a unique opportunity to assess progress, seasonality, and also introduce some new analyses and visualizations. With this we anticipate new levels of insights
for us to understand progress on this important issue.

Increased Measurement Disclosures First, and Rigor Next:

Global transparency reporting is a significant endeavor. Two of the principle authorized metrics in the report on consumer safety and advertiser safety largely rely on a global sample measurement methodology. Under these global metrics, samples and forecasts are made for local markets and languages. We will work with platforms and local markets to disclose these local samples, and also explore ways to report on key actions at the local level.

Executive Summary

YouTube

Our Commitment to Responsibility

At YouTube, we work hard to maintain a safe and vibrant community. **Responsibility remains our #1 priority, and we continue to approach this work from several angles, 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 clear **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.

Over the past several years, machine learning has transformed our ability to tackle how we remove violative content at scale. Because of our ongoing investments in machine learning, **in H1 2022 we were able to detect >92% 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.

In addition to our Community Guidelines, we enforce a second set of policies, our **Ad Friendly Guidelines,** which set the standard for which videos are eligible for ads. These guidelines are more restrictive than our Community Guidelines and adhere to the GARM brand safety floor. We measure our effectiveness of enforcing these guidelines through our Advertiser Safety Error Rate, included in this report as a GARM Authorized metric. As part of our Media Rating Council (MRC) accreditation for content-level brand safety, we also publish our effectiveness at enforcing Ad Friendly Guidelines here.

Additionally, we continue to make investments in other areas of critical importance, like transparency.

YouTube Community Guidelines Enforcement Report
Every quarter our Community Guidelines Enforcement Report showcases data demonstrating the vast impact of our enforcement work and the progress we've made with regards to content on the platform that violates our Community Guideline policies. This includes flagging (human and automated), video, channel, and comment removals, appeals and reinstatements, and highlighted policy verticals.

Since the first report launched in April 2018, we have updated the data on a quarterly basis and, like other Transparency Reports we offer at Google, the data we share—and the way we share it—evolves over time. Most recently, in Q2 2022 we updated our Community Guidelines Enforcement report to include the number of videos violating our misinformation policies (e.g. medical & general misinformation). Previously, this data was disclosed in the "Spam, Misleading, and Scams" vertical.

Our most recent YouTube transparency report, and past quarterly editions, can be reviewed here. **For the purposes of the GARM Aggregated Measurement Report Volume 4, we have included quarterly data and critical insights from our last four transparency reports in this resource, represented as bi-annual aggregations (1H 2022 & 2H 2021).**

Academic Research
- **YouTube's Violative View Rate:** In September 2021, Professor Arnold Barnett at MIT Sloan published a report to assess the completeness and the appropriateness of our Violative View Rate (VVR) metric. Bartnett found the methodology for VVR statistically sound and saw VVR as an accurate way to estimate the effectiveness of our enforcement of Community Guidelines.
- **YouTube Researcher Program:** In July 2022, YouTube announced the YouTube Researcher Program to provide qualified academic researchers from around the world with scaled access to YouTube's Data API and robust in-house technical support for academic research projects. The program is still in its early stages, and we plan to build out additional features based on feedback from the research community over time.

MRC Content-level Brand Safety Accreditation
In April 2022, **the MRC re-accredited YouTube for content-level Brand Safety**, making YouTube the industry's first digital platform to receive such annual accreditation *and* to receive it two years in a row. YouTube's continued MRC 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 (included in this GARM resource as an Authorized metric) as an MRC accredited metric for the first time.

This applies to YouTube in-stream video inventory purchased through Google Ads, Display & Video 360, and YouTube Reserve services, excluding video discovery, YouTube Kids, and Live Stream.

YouTube

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's Chief Product Officer 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 4,** enforcement **of our misinformation policies is represented as part of YouTube's "Spam, deceptive practices, scams and misinformation" category.**

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 4, our 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.**

July through December 2021
Between July and December 2021, YouTube removed over 9.9 million videos for violating Community Guidelines. **The vast majority (>93%) of these videos were first flagged by machines rather than humans. Over 457k video removals were appealed, and we reinstated <133k of those videos.** YouTube terminated over 8.6 million channels for violating our Community Guidelines, the overwhelming majority which were terminated for violating our spam policies. Our violative view rate (VVR) ranged from 0.09-0.11% in Q3 2021 to 0.12-0.14% in Q4 2021. This means that out of every 10,000 views on YouTube in Q4 only 12-14 came from violative content. Lastly, YouTube removed more than 2 billion comments, the majority of which were spam; 99% of removed comments were detected automatically.

In H2'21, we also made progress on one of the industry's most challenging but critical areas: **misinformation.**
- In September 2021, we further strengthened our longstanding medical misinformation Community Guidelines to remove content with false claims about currently-administered vaccines that are approved and confirmed to be safe and effective by local health authorities and WHO.
- In October 2021, we expanded our Advertiser-friendly Guidelines to address Climate Change misinformation, prohibiting the monetization of content that contradicts well-established scientific consensus around the existence and causes of climate change. We are committed to combating misinformation as we evolve our policies in tandem with subject experts.

YouTube

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)

	Latest Period		Previous Period	
GARM Metric	**Q2 2022**	**Q1 2022**	**Q4 2021**	**Q3 2021**
Violative View Rate	0.09-0.11%	0.09-0.11%	0.12-0.14%	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

	Latest Period		Previous Period	
GARM Metric	**Q2 2022**	**Q1 2022**	**Q4 2021**	**Q3 2021**
Advertiser 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[1]		Actors Actioned[2]		Comments Actioned	
	Latest Period Q1 & Q2 2022	Previous Period Q3 & Q4 2021	Latest Period Q1 & Q2 2022	Previous Period Q3 & Q4 2021	Latest Period Q1 & Q2 2022	Previous Period Q3 & Q4 2021
Nudity or sexual	1,320,730	1,839,103	335,801	313,035	2,365,605	4,908,817
Child safety	2,351,206	3,168,476	160,614	95,007	194,720,777	299,902,525
Harmful or dangerous	1,015,223	585,755	61,847	36,201	118,713	17,713
Promotion of violence and violent extremism	133,711	322,751	19,890	21,803	446,506	291,379
Harassment and cyberbullying	922,039	606,582	119,999	129,778	255,959,014	292,998,647
Violent or graphic	1,724,913	2,202,748	18,021	27,916	44,356	320,893
Spam, deceptive practices, scams and misinformation	641,280	947,384	7,538,816	7,892,531	1,121,879,591	1,669,336,620
Hateful or abusive	241,635	203,532	75,559	78,844	122,379,812	94,132,595
Impersonation	n/a	n/a	47,741	51,791	n/a	n/a
Other	28,880	107,766	19,570	9,411	200	31,123

[1] Content Actioned for YouTube is Videos Removed
[2] Actors Actioned for YouTube is Channels Removed

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[1]		Actors Actioned[2]		Comments Actioned	
	Latest Period Q1 & Q2 2022	Previous Period Q3 & Q4 2021	Latest Period Q1 & Q2 2022	Previous Period Q3 & Q4 2021	Latest Period Q1 & Q2 2022	Previous Period Q3 & Q4 2021
Nudity or sexual	1,320,730	1,839,103	335,801	313,035	2,365,605	4,908,817
Child safety	2,351,206	3,168,476	160,614	95,007	194,720,777	299,902,525
Harmful or dangerous	1,015,223	585,755	61,847	36,201	118,713	17,713
Promotion of violence and violent extremism	133,711	322,751	19,890	21,803	446,506	291,379
Harassment and cyberbullying	922,039	606,582	119,999	129,778	255,959,014	292,998,647
Violent or graphic	1,724,913	2,202,748	18,021	27,916	44,356	320,893
Spam, deceptive practices, scams and misinformation	641,280	947,384	7,538,816	7,892,531	1,121,879,591	1,669,336,620
Hateful or abusive	241,635	203,532	75,559	78,844	122,379,812	94,132,595
Impersonation	n/a	n/a	47,741	51,791	n/a	n/a
Other	28,880	107,766	19,570	9,411	200	31,123

[1] Content Actioned for YouTube is Videos Removed
[2] 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** Q1 & Q2 2022	**Previous Period** Q3 & Q4 2021
Total Video Removals	8,379,617	9,984,097
Removed videos by views: 0 views	32.8%	34.8%
Removed videos by views: 1-10	35.1%	36.6%
Removed videos by views: 10+	32.1%	28.6%

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** Q1 & Q2 2022	**Previous Period** Q3 & Q4 2021
Content Appealed: Videos	443,507	457,426
Content Reinstated: Video	59,475	132,737

Global Alliance for Responsible Media

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.

GARM Brand Safety Floor Category + Definition • Defines content that can monetize. • Aligned with YouTube's Ad Friendly Guidelines, a higher bar than Community Guidelines.	**Relevant YouTube Community Guidelines** ● Governs content that can live on YouTube. ● Our Community Guidelines Enforcement Report measures our enforcement of these guidelines.
Adult & Explicit Sexual Content • Illegal sale, distribution, and consumption of child pornography • Explicit or gratuitous depiction of sexual acts, and/or display of genitals, real or animated	**Nudity and sexual Content** 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. **Child safety** 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.
Arms & Ammunition • Promotion and advocacy of Sales of illegal arms, rifles, and handguns • Instructive content on how to obtain, make, distribute, or use illegal arms • Glamorization of illegal arms for the purpose of harm to others • Use of illegal arms in unregulated environments	**Firearms** 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.
Crime & Harmful acts to individuals and Society, Human Right Violations • 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.) • Harassment of bullying of individuals and groups	**Harmful or dangerous Content** YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death. **Hate speech** 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. **Harassment and cyberbullying** 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 group status or physical traits.
Death, Injury or Military Conflict • Promotion, incitement or advocacy of violence, death or injury • Murder or willful bodily harm to others • Graphic depictions of willful harm to others • Incendiary content provoking, enticing, or evoking military aggression • Live action footage/photos of military actions & genocide or other war crimes	**Violent or graphic content** Violent or gory content intended to shock or disgust viewers, or content encouraging others to commit violent acts are not allowed on YouTube. **Harmful or dangerous content** YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death. **Suicide & self-injury** We do not allow content on YouTube that promotes suicide, self-harm, or is intended to shock or disgust users.

Global Alliance for Responsible Media

Online piracy • Pirating, Copyright infringement, & Counterfeiting	**Fake engagement** 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.
	Impersonation 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.
	Sale of illegal or regulated goods or services Content intended to sell certain regulated goods and services is not allowed on YouTube. Such as: Counterfeit documents or currency
	YouTube's Terms of Service Also covered in YouTube's Terms of Service
Hate speech & acts of aggression • 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	**Hate speech** 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.
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust • Excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult.	**Violent or graphic content** Violent or gory content intended to shock or disgust viewers, or content encouraging others to commit violent acts are not allowed on YouTube.
	Age restriction 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.
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol • 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 • Promotion and advocacy of Tobacco and e-cigarette (Vaping) & Alcohol use to minors	**Sale of illegal or regulated goods or services** 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.
	Harmful or dangerous content YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death.
Spam or Harmful Content • Malware/Phishing	**Spam deceptive practices, scams and misinformation** 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.
Misinformation • The presence of verifiably false or willfully misleading content that is directly connected to user or societal harm	
Terrorism • Promotion and advocacy of graphic terrorist activity involving defamation, physical and/or emotional harm of individuals, communities, and society	**Violent criminal organizations** 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.
Debated Sensitive Social Issue • Insensitive, irresponsible and harmful treatment of debated social issues and related acts that demean a particular group or incite great conflict	
Other	**Other** 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.

Meta

Our Commitment to Responsibility and Transparency on Facebook and Instagram

We want Facebook and Instagram to be places where people have a voice. 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 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 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 continue to build technologies like RIO, WPIE and XLM-R that help us identify harmful content faster, across languages and different content types. These efforts, our continued focus on AI research help 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.

We reduce prevalence of violating content in a number of ways, including improvements in detection and enforcement and reducing problematic content in Feed. These tactics have enabled us to cut hate speech prevalence by more than half within the last year on Facebook, and we're using these same tactics across policy areas like violence and incitement and bullying and harassment. 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 new cross-problem AI system to consolidate learnings for all three to better address each violation area.

We're also using warning screens to educate and discourage people from posting something that may include hostile speech such as bullying and harassment violating our Community Standards. The screens appear after someone has typed a post or comment explaining that the content may violate our rules and may be hidden or distribution reduced. Repeatedly posting this content could result in an account being disabled or deleted. We see these are working, too. In a one-week period, about 50% of the time the comment was edited or deleted by the user based on these warnings.

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, 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. We're adding Univision as another US partner to cover Spanish language content.

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 and partners can be monetized - 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 behaviour that may result in advertising restrictions being placed on a business account or its assets (an ad account, Page or user account).

General trends over the year on Facebook and Instagram

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, content actioned (and the percentage of that we found proactively), and content appealed and restored. Since 2020 we have released the Community Standards Enforcement Report every quarter. Our Q4 2021 report was our 12th report and some of our long-term trends include:

- Prevalence of bullying & harassment on Facebook has continued to decrease, in Q2 2022 it was 0.08% down from 0.14%-0.15% when we first began reporting it.
- Our proactive rate (the percentage of content we took action on that we found before a user reported it to us) is over 90% for 13 out of 14 policy areas on Facebook and 11 out of 12 on Instagram.
- Our ongoing commitment and investments in AI have enabled us to show improvements across both content actioned and proactive rate across many policy areas.
- While our technology for identifying and removing violating content is improving, there will continue to be areas where we rely on people to both review content and train our technology.

Meta

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 develop new features and tools so people can foster their relationships in a safe, supportive environment.

To keep young people safe:
• We set teens' accounts to private when they join Instagram.
• We prevent adults they don't follow from sending them DMs.
• 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' 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.

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.

Overall trends Q2 2022 (our latest report) on Facebook and Instagram

Prevalence of harmful content on Facebook and Instagram remained relatively consistent or decreased from Q1 2022 to Q2 2022 across most of our policy areas, meaning the vast majority of content that users encounter does not violate our standards.

Bullying and harassment prevalence has decreased on both Facebook and Instagram in the last quarter.
• On Facebook, prevalence was 0.08%, or 8 views of content per 10,000 views in Q2 2022, down from 0.14-0.15% when we first began reporting it in Q3 2021.
• On Instagram, prevalence was 0.04%, or 4 views of content per 10,000 views in Q2 2022, down from 0.5% when we first began reporting it in Q3 2021.

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

Meta

2022 Roadmap

We plan to continue to build on and improve our reports as our goal continues to be to lead the technology industry in transparency, and we'll continue to share more metrics as part of this effort. We're committed to sharing meaningful data and reporting so that we can be held accountable for our progress, even when it shows areas where we need to do better. In order to report an metric externally we go through a very thorough process to ensure confidence in our metrics before releasing them publicly.

First Party Content Based Controls for Feed
Across Meta, we are designing suitability controls to give advertisers control over where their ads are shown. In November, 2021 we announced our commitment to build first party pre-campaign content-based suitability controls for Facebook and Instagram Feeds. We have been working closely with GARM as we develop these controls, which will be aligned with the GARM Suitability Framework.

We have begun scoping and building these new controls for Facebook and Instagram Feeds focused on primarily English speaking markets, and early testing has begun as of Q3 2022 with advertisers. We expect these controls to be broadly available in early 2023. These controls will be subject to an MRC audit once they have been fully built and are more operational. Over the course of 2023, we will expand placement coverage to include Stories, Reels, Video Feeds, Instagram Explore and other surfaces across Facebook and Instagram, as well as expanding to additional languages.

Third Party Brand Suitability Verification in Feed
After an extensive vetting process, we've selected Zefr as the initial partner for providing independent reporting on the context in which ads appear on Facebook Feed. We have worked together to develop a 3rd party post campaign solution to measure and verify the suitability of adjacent content to ads in Facebook Feed, and have started small scale testing in the third quarter of 2022. We expect this solution to be broadly available in early 2023.

Independent Verification

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.

Community Standards Enforcement Report audit
Last year, we asked EY (Ernst & Young) to verify the metrics of our Community Standards Enforcement Report since we don't believe we should grade our own homework. EY's independent assessment focused on the metrics we report in the Community Standards Enforcement Report. EY found that the calculation of the metrics in our 2021 Q4 Community Standards Enforcement Report were fairly stated, and our internal controls are suitably designed and operating effectively. This assessment builds on work we started in 2018. Read more detail here.

MRC (Media Ratings Council) audit
In August 2020, we committed to undertaking and releasing independent, third party assessments of our brand safety and suitability solutions. We're thrilled to announce that we've received the first wave of content-level Brand Safety accreditation from the MRC on Facebook.

We also recently announced that testing started for our content-based inventory filter for Facebook Feed and Instagram Feed—we will continue to iterate on the brand safety and suitability solutions we make available to businesses, and plan to extend the MRC audit to the content-based inventory filter controls for Feed once they are generally available.

Meta

Transparency Reporting and Methodology on Facebook and Instagram

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 our Widely Viewed Content Report

Our Community Standards Enforcement Report measures:

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 two types of sampling (stratified and random) 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., 0.08%-0.10%)
- 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 and are continually working to expand coverage of the metric.

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.

Meta

Mapping of GARM Brand Safety Floor to Facebook Community Standards

GARM/4As Category	Facebook Policy
Adult and Explicit Sexual Content	Adult Nudity and Sexual Activity, Child Sexual Exploitation, Abuse and Nudity, Sexual Solicitation
Arms and Ammunition	Violence and Incitement, Coordinating Harm and Publicizing Crime, Restricted Goods and Services
Crime and Harmful Acts to Individuals and Society and Human Right Violations	Adult Nudity and Sexual Activity , Violence and Incitement, Bullying and Harassment Violent and Graphic Content, Child Sexual Exploitation, Abuse and Nudity; Suicide and Self-Injury, Cruel and Insensitive; Human Exploitation, Dangerous Individuals and Organizations , Coordinating Harm and Publicizing Crime, Restricted Goods and Services, Fraud and Deception
Death, Injury or Military Conflict	Violence and Incitement, Violent and Graphic Content, Cruel and Insensitive, Suicide and Self-Injury
Online Piracy	Intellectual Property Fraud and Deception
Hate Speech and Acts of Aggression	Hate Speech, Bullying and Harrassment, Dangerous Individuals and Organizations
Obscenity and Profanity, including language, gestures and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech, Bullying and Harrassment
Illegal Drugs/Tobacco/E-cigarettes/ Vaping/Alcohol	Restricted Goods and Services
Spam or Harmful Content	Cybersecurity Spam
Terrorism	Dangerous Individuals and Organizations
Debated Sensitive Social Issues	Hate Speech, Bullying and Harrassment
Misinformation	Misinformation

Additional policies not covered	Facebook Policy
Floor focuses online and not on offline/real-world fraud Floor does not include census and voter interference/fraud Floor does not include coverage for creepshots	Fraud and Deception Coordinating Harm and Publicizing Crime Adult Sexual Exploitation

Other Facebook Policies Floor does not address

Privacy Violations

Account Integrity and Authentic Integrity

Inauthentic Behavior

Memorialization

User Requests

Additional Protections for Minors

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.

		Latest Period		Previous Period		
GARM Category	Relevant Policy	Q2 2022	Q1 2022	Q4 2021	Q3 2021	Commentary
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	**0.04%**	**0.04%**	**0.03%**	**0.02-0.03%**	Adult Nudity and Sexual Activity: Prevalence increased from Q4 2021 to Q1 2022 due to an increase in spam actors sharing large volumes of videos containing nudity.

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: 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	Regulated Goods: Firearms	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	Violence and Incitement: Prevalence was 0.03% in Q2 2022 due to a decrease in overall and violating comments. Prevalence in Q1 2022 decreased due to the improvement and expansion of our proactive detection technology.
	Violence and Incitement	**0.03%**	**0.03%**	**0.03-0.04%**	**0.04-0.05%**	

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
		Q2 2022	Q1 2022	Q4 2021	Q3 2021	
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	0.04%	0.04%	0.03%	0.02-0.03%	
	Violence and Incitement	0.03%	0.03%	0.03-0.04%	0.04-0.05%	
	Violent and Graphic Content	0.04%	0.03-0.04%	0.03-0.04%	0.04%	
	Bullying and Harassment	0.08-0.09%	0.09-0.10%	0.11-0.12%	0.14-0.15%	
	Child Nudity and Sexual Exploitation	N/A	N/A	N/A	N/A	Adult Nudity and Sexual Activity: Prevalence increased from Q4 2021 to Q1 2022 due to an increase in spam actors sharing large volumes of videos containing nudity.

Violence and Incitement: Prevalence was 0.03% in Q2 2022, which marks a decrease from Q3 2021, when we first reported prevalence, due to a decrease in overall and violating comments.

Violent and Graphic Content: Prevalence has remained relatively consistent.

Bullying and Harassment: Prevalence in Q2 2022 decreased due to AI improvements. Prevalence decreased in Q1 2022 due to the continued impact of refinements we made to our policies to identify this content in mid Q4.

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: 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%	
	Regulated Goods: 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 Facebook and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q2 2022	Q1 2022	Q4 2021	Q3 2021	
Death, Injury or Military Conflict	Violent and Graphic Content	0.04%	0.03-0.04%	0.03-0.04%	0.04%	Violent and Graphic Content: Prevalence has remained relatively consistent. Violence and Incitement: Prevalence was 0.03% in Q2 2022 due to a decrease in overall and violating comments. Prevalence in Q1 2022 decreased due to the improvement and expansion of our proactive detection technology.
	Violence and Incitement	0.03%	0.03%	0.03-0.04%	0.04-0.05%	
	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 or 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%	0.02-0.03%	0.03%	Hate Speech: Prevalence has remained relatively consistent. Bullying and Harassment: Prevalence in Q2 2022 decreased due to AI improvements. Prevalence decreased in Q1 2022 due to the continued impact of refinements we made to our policies to identify this content in mid Q4. 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	0.08-0.09%	0.09-0.10%	0.11-0.12%	0.14-0.15%	

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
		Q2 2022	Q1 2022	Q4 2021	Q3 2021	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	0.02%	0.02%	0.02-0.03%	0.03%	Hate Speech: Prevalence has remained relatively consistent. Bullying and Harassment: Prevalence in Q2 2022 decreased due to AI improvements. Prevalence decreased in Q1 2022 due to the continued impact of refinements we made to our policies to identify this content in mid Q4.
	Bullying and Harassment	0.08-0.09%	0.09-0.10%	0.11-0.12%	0.14-0.15%	
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: 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%	0.02-0.03%	0.03%	Hate Speech: Prevalence has remained relatively consistent. Bullying and Harassment: Prevalence in Q2 2022 decreased due to AI improvements. Prevalence decreased in Q1 2022 due to the continued impact of refinements we made to our policies to identify this content in mid Q4.
	Bullying and Harassment	0.08-0.09%	0.09-0.10%	0.11-0.12%	0.14-0.15%	

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

		Latest Period						Previous Period	
GARM Category	Relevant Policy	Q2 2022		Q1 2022		Q4 2021		Q3 2021	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	38.4m	97.2%	31m	96.7%	27.3m	97.7%	34.7m	98.8%
	Child Endangerment: Nudity and Physical Abuse	1.9m	97.3%	2.1m	97.8%	1.8m	97.5%	1.8m	97.1%
	Child Endangerment: Sexual Exploitation	20.4m	99.1%	16.5m	96.4%	19.8m	99.0%	21.2m	99.1%
Arms & Ammunition	Regulated Goods: Firearms	1.6m	94.4%	1.2m	94.6%	1.5m	92.0%	1.1m	94.1%
	Violence and Incitement	19.3m	98.2%	21.7m	98.1%	12.4m	96.6%	13.6m	96.7%
Crime & Harmful acts to Individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	38.4m	97.2%	31m	96.7%	27.3m	97.7%	34.7m	98.8%
	Violence and Incitement	19.3m	98.2%	21.7m	98.1%	12.4m	96.6%	13.6m	96.7%
	Violent and Graphic Content	45.9m	99.5%	26.1m	99.5%	25.2m	99.5%	26.6m	99.4%
	Bullying and Harassment	8.2m	76.8%	9.5m	67.0%	8.2m	58.8%	9.2m	59.4%
	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	1.9m	97.3%	2.1m	97.8%	1.8m	97.5%	1.8m	97.1%
	Child Endangerment: Sexual Exploitation	20.4m	99.1%	16.5m	96.4%	19.8m	99.0%	21.2m	99.1%
	Suicide and Self-Injury	11.3m	99.1%	6.8m	98.8%	6.1m	98.8%	8.5m	99.0%
	Regulated Goods: Firearms	1.6m	94.4%	1.2m	94.6%	1.5m	92%	1.1m	94.1%
Death, Injury or Military Conflict	Violent and Graphic Content	45.9m	99.5%	26.1m	99.5%	25.2m	99.5%	26.6m	99.4%
	Violence and Incitement	19.3m	98.2%	21.7m	98.1%	12.4m	96.6%	13.6m	96.7%
	Suicide and Self Injury	11.3m	99.1%	6.8m	98.8%	6.1m	98.8%	8.5m	99%
Online Piracy	Intellectual Property: Copyright	N/A*	N/A*	N/A*	N/A*	N/A*	N/A*	N/A*	N/A*
	Intellectual Property: Counterfeit	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*

* We release the H1 2022 figures in Nov, 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 Facebook

GARM Category	Relevant Policy	Q2 2022		Q1 2022		Q4 2021		Q3 2021	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
Hate speech & acts of aggression	Hate Speech	13.5m	95.6%	15.1m	95.6%	17.4m	95.9%	22.3m	96.5%
	Dangerous Organizations: Terrorism	13.6m	98.9%	16.1m	98.8%	7.7m	97.7%	10.6m	97.9%
	Dangerous Organizations: Organized Hate	2.3m	96.9%	2.5m	96.9%	1.6m	96.1%	2m	96.4%
	Bullying and Harassment	8.2m	76.8%	9.5m	67.0%	8.2m	58.8%	9.2m	59.4%
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	13.5m	95.6%	15.1m	95.6%	17.4m	95.9%	22.3m	96.5%
	Bullying and Harassment	8.2m	76.8%	9.5m	67.0%	8.2m	58.8%	9.2m	59.4%
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	3.9m	98.1%	3.3m	97.7%	4m	97.9%	2.7m	96.7%
Spam or Harmful Content	Spam	734.2m	99.2%	1.8b	99.7%	1.2b	99.6%	777.2m	99.6%
Terrorism	Dangerous Organizations: Terrorism	13.6m	98.9%	16.1m	98.8%	7.7m	97.7%	10.6m	97.9%
	Dangerous Organizations: Organized Hate	2.3m	96.9%	2.5m	96.9%	1.6m	96.1%	2m	96.4%
Debated Sensitive Social Issue	Hate Speech	13.5m	95.6%	15.1m	95.6%	17.4m	95.9%	22.3m	96.5%
	Bullying and Harassment	8.2m	76.8%	9.5m	67.0%	8.2m	58.8%	9.2m	59.4%

Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Facebook

GARM Category	Relevant Policy	Commentary
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	Adult Nudity and Sexual Activity: Content actioned increased from 31 million in Q1 2022 to 38.4 million in Q2 2022 due to the continued impact of an increase in spam actors sharing large volumes of videos containing nudity that violated our policy.
	Child Endangerment: Nudity and Physical Abuse	Child Endangerment: Nudity and Physical Abuse: Content actioned for child nudity and physical abuse decreased from 2.3 million pieces of content in Q2 2021 to 1.8 million in Q3 2021.
	Child Endangerment: Sexual Exploitation	Child Endangerment: Sexual Exploitation: Content actioned increased from 16.5 million in Q1 2022 to 20.4 million in Q2 2022 due to an increase in enforcement on viral content and old, violating content detected by our media-matching technology. Proactive rate increased from 96.4% in Q1 2022 to 99.1% in Q2 2022, returning to pre-Q1 levels after a spike in reported viral links in February. Content actioned decreased from 19.8 million in Q4 2021 to 16.5 million in Q1 2022 due to fewer actions taken on old, violating content than in Q4. Proactive rate decreased from 99% in Q4 2021 to 96.4% in Q1 2022 due to a spike in reported viral links in February.
Arms & Ammunition	Regulated Goods: Firearms	Regulated Goods: Firearms: Content actioned increased from 1.2 million in Q1 2022 to 1.6 million in Q2 2022 due to improvements made to our proactive detection technology. Content actioned increased from 1.1 million in Q3 2021 to 1.5 million in Q4 2021 due to improved and expanded proactive detection technologies. Content actioned decreased from 1.9 million pieces of content in Q1 2021 to 1.5 million in Q2 2021 due to a decline in violating firearms content.
	Violence & Incitement	Violence & Incitement: Content actioned increased from 12.4 million in Q4 2021 to 21.7 million in Q1 2022 due to the improvement and expansion of our proactive detection technology.
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	Adult Nudity and Sexual Activity: Content actioned increased from 31 million in Q1 2022 to 38.4 million in Q2 2022 due to the continued impact of an increase in spam actors sharing large volumes of videos containing nudity that violated our policy.
	Violence & Incitement	Violence & Incitement: Content actioned increased from 12.4 million in Q4 2021 to 21.7 million in Q1 2022 due to the improvement and expansion of our proactive detection technology.
	Violent and Graphic Content	Violent and Graphic Content: Content actioned increased from 26.1 million in Q1 2022 to 45.9 million in Q2 2022 due to improvements made to our media-matching technology that had enabled us to take down more old, violating content. In Q1, we adjusted our media-matching technology and were able to take action on old, violating content, and we also made improvements to our proactive detection technology to detect and remove more videos automatically.
	Bullying and Harassment	Bullying and Harassment: Proactive rate increased from 67% in Q1 2022 to 76.8% in Q2 2022 due to expanding our improved prioritization for proactive detection technology to more languages. Proactive rate increased from 58.8% in Q4 2021 to 67% in Q1 2022 due to the improvement and expansion of our proactive detection technology.
	Child Endangerment: Nudity and Physical Abuse	Child Endangerment: Nudity and Physical Abuse: Content actioned for child nudity and physical abuse decreased from 2.3 million pieces of content in Q2 2021 to 1.8 million in Q3 2021.
	Child Endangerment: Sexual Exploitation	Child Endangerment: Sexual Exploitation: Content actioned increased from 16.5 million in Q1 2022 to 20.4 million in Q2 2022 due to an increase in enforcement on viral content and old, violating content detected by our media-matching technology. Proactive rate increased from 96.4% in Q1 2022 to 99.1% in Q2 2022, returning to pre-Q1 levels after a spike in reported viral links in February. Content actioned decreased from 19.8 million in Q4 2021 to 16.5 million in Q1 2022 due to fewer actions taken on old, violating content than in Q4. Proactive rate decreased from 99% in Q4 2021 to 96.4% in Q1 2022 due to a spike in reported viral links in February.
	Suicide and Self-Injury	Suicide and Self Injury: Content actioned increased from 6.8 million in Q1 2022 to 11.3 million in Q2 2022 primarily due to improvements made to our media matching technology. Content actioned decreased from 8.5 million in Q3 2021 to 6.1 million in Q4 2021, continuing the return to pre-Q2 levels, when we resolved a technical issue which enabled us to remove old, violating content detected by our media-matching technology. Content actioned decreased from 16.8 million to 8.5 million after the increase in Q2 2021 where we resolved a technical issue, which enabled us to remove old, violating content detected by our media-matching technology. Content actioned increased from 5.2 million pieces of content in Q1 2021 to 16.8 million in Q2 2021. This was due to us resolving a technical issue, which enabled us to remove old, violating content detected by our media-matching technology. In Q1, we adjusted our media-matching technology and were able to take action on old, violating content. A technical issue in Q1 also caused our detection technology to take action on some older content that wasn't violating.
	Regulated Goods: Firearms	Regulated Goods: Firearms: Content actioned decreased from 1.5 million in Q2 2021 to 1.1 million in Q3 2021 going back to pre-2021 levels. Content actioned increased from 1.1 million in Q3 2021 to 1.5 million in Q4 2021 due to improved and expanded proactive detection technologies. Content actioned decreased from 1.9 million pieces of content in Q1 2021 to 1.5 million in Q2 2021 due to a decline in violating firearms content.
Death, Injury or Military Conflict	Violent and Graphic Content	Violent and Graphic Content: Content actioned increased from 26.1 million in Q1 2022 to 45.9 million in Q2 2022 due to improvements made to our media-matching technology that had enabled us to take down more old, violating content. In Q1, we adjusted our media-matching technology and were able to take action on old, violating content, and we also made improvements to our proactive detection technology to detect and remove more videos automatically.
	Violence and Incitement	Violence & Incitement: Content actioned increased from 12.4 million in Q4 2021 to 21.7 million in Q1 2022 due to the improvement and expansion of our proactive detection technology.
	Suicide and Self Injury	Suicide and Self Injury: Content actioned increased from 6.8 million in Q1 2022 to 11.3 million in Q2 2022 primarily due to improvements made to our media matching technology. Content actioned decreased from 8.5 million in Q3 2021 to 6.1 million in Q4 2021, continuing the return to pre-Q2 levels, when we resolved a technical issue which enabled us to remove old, violating content detected by our media-matching technology. Content actioned decreased from 16.8 million to 8.5 million after the increase in Q2 2021 where we resolved a technical issue, which enabled us to remove old, violating content detected by our media-matching technology. Content actioned increased from 5.2 million pieces of content in Q1 2021 to 16.8 million in Q2 2021. This was due to us resolving a technical issue, which enabled us to remove old, violating content detected by our media-matching technology. In Q1, we adjusted our media-matching technology and were able to take action on old, violating content. A technical issue in Q1 also caused our detection technology to take action on some older content that wasn't violating.

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	We report this metric monthly in a 6 month report. Our current report has data for January - June 2021. 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.

Our proactive rate figure here constitutes the volume of content removed in response to an IP report relative to the volume of content reported, reflected as a percentage. In prior transparency reports, the Removal Rate constituted the percentage of total IP reports that resulted in some or all reported content being removed. Beginning in the July 2019 reporting period, we have adjusted the way we calculate Removal Rate to reflect the percentage of reported content removed, rather than the percentage of reports resulting in removals. Because a single IP report can identify multiple pieces of content, this figure offers a more complete picture of the total content removed from the platform based on an IP report. |
	Intellectual Property: Counterfeit	
	Intellectual Property: Trademark	
Hate speech & acts of aggression	Hate Speech	Hate Speech: Content actioned increased from 25.2 million pieces of content in Q1 2021 to 31.5 million in Q2 2021, driven by improvements to our proactive detection technology in late Q1. Content actioned decreased from 22.3 million in Q3 2021 to 17.4 million in Q4 2021 due to a decrease in overall and violating comments. Content actioned decreased from 31.5 million in Q2 2021 to 22.3 million in Q3 2021 due to adjustments to our proactive detection technology to improve precision of our enforcement actions.
	Dangerous Organizations: Terrorism	Dangerous Organizations: Terrorism: Content actioned decreased from 16.1 million in Q1 2022 to 13.6 million in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Content actioned increased in Q1 2022 from Q4 2021 for both Terrorism and Organized Hate due to an increase in enforcement on non-violating content due to a bug in our media-matching technology that was later fixed and the content was restored. Content actioned decreased from 10.6 million pieces of content in Q3 2021 to 7.7 million in Q4 2021, returning back to pre-Q3 levels following an update to our media-matching technology that enabled us to take down more old content.
	Dangerous Organizations: Organized Hate	Dangerous Organizations: Organized Hate: Content actioned increased in Q1 2022 from Q4 2021 for both Terrorism and Organized Hate due to an increase in enforcement on non-violating content due to a bug in our media-matching technology that was later fixed and the content was restored. Content actioned decreased from 2 million in Q3 2021 to 1.6 million in Q4 2021 as a continuation from the updates we made to our media matching technology in Q3 to improve the precision of our decisions.
	Bullying and Harassment	Bullying and Harassment: Proactive rate increased from 67% in Q1 2022 to 76.8% in Q2 2022 due to expanding our improved prioritization for proactive detection technology to more languages. Proactive rate increased from 58.8% in Q4 2021 to 67% in Q1 2022 due to the improvement and expansion of our proactive detection technology.
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	Hate Speech: Content actioned increased from 25.2 million pieces of content in Q1 2021 to 31.5 million in Q2 2021, driven by improvements to our proactive detection technology in late Q1. Content actioned decreased from 22.3 million in Q3 2021 to 17.4 million in Q4 2021 due to a decrease in overall and violating comments. Content actioned decreased from 31.5 million in Q2 2021 to 22.3 million in Q3 2021 due to adjustments to our proactive detection technology to improve precision of our enforcement actions.
	Bullying and Harassment	Bullying and Harassment: Proactive rate increased from 67% in Q1 2022 to 76.8% in Q2 2022 due to expanding our improved prioritization for proactive detection technology to more languages. Proactive rate increased from 58.8% in Q4 2021 to 67% in Q1 2022 due to the improvement and expansion of our proactive detection technology.
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	Regulated Goods: Drugs: Content actioned increased from 2.7 million in Q3 2021 to 4 million in Q4 2021 due to improvements made to our proactive detection technologies. Content actioned increased from 2.3 million in Q2 2021 to 2.7 million in Q3 2021 due to improved proactive detection technologies we launched in late Q2 2021. Content actioned decreased from 3.2 million pieces of content in Q1 2021 to 2.3 million in Q2 2021. In Q1, we adjusted our proactive detection technology to continue improving precision, which resulted in fewer content removals. Content actioned decreased in Q1 2021, after making adjustments to our automation in order to improve accuracy, this temporarily decreased the amount of content we took action on in Q1.
Spam or Harmful Content	Spam	Spam: Content actioned decreased from 1.8 billion in Q1 2022 to 734.2 million in Q2 2022 returning to pre-Q4 levels as we began rate limiting users who made a large volume of posts in a short duration of time. Instead of deleting all of the content after being posted, we now prevent users from posting a large volume of content in a short span of time. Content actioned increased from 1.2 billion in Q4 2021 to 1.8 billion in Q1 2022 due to actions on a small number of users making a large volume of posts. Content actioned increased from 777.2 million in Q3 2021 to 1.2 billion in Q4 2021 due to a large amount of violating content removed in December. Fluctuations in enforcement metrics for Spam are expected due to the highly adversarial nature of this space.
Terrorism	Dangerous Organizations: Terrorism	Dangerous Organizations: Terrorism: Content actioned decreased from 16.1 million in Q1 2022 to 13.6 million in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Content actioned increased in Q1 2022 from Q4 2021 for both Terrorism and Organized Hate due to an increase in enforcement on non-violating content due to a bug in our media-matching technology that was later fixed and the content was restored. Content actioned decreased from 10.6 million pieces of content in Q3 2021 to 7.7 million in Q4 2021, returning back to pre-Q3 levels following an update to our media-matching technology that enabled us to take down more old content.
	Dangerous Organizations: Organized Hate	Dangerous Organizations: Organized Hate: Content actioned increased in Q1 2022 from Q4 2021 for both Terrorism and Organized Hate due to an increase in enforcement on non-violating content due to a bug in our media-matching technology that was later fixed and the content was restored. Content actioned decreased from 2 million in Q3 2021 to 1.6 million in Q4 2021 as a continuation from the updates we made to our media matching technology in Q3 to improve the precision of our decisions.
Debated Sensitive Social Issue	Hate Speech	Hate Speech: Content actioned increased from 25.2 million pieces of content in Q1 2021 to 31.5 million in Q2 2021, driven by improvements to our proactive detection technology in late Q1. Content actioned decreased from 22.3 million in Q3 2021 to 17.4 million in Q4 2021 due to a decrease in overall and violating comments. Content actioned decreased from 31.5 million in Q2 2021 to 22.3 million in Q3 2021 due to adjustments to our proactive detection technology to improve precision of our enforcement actions.
	Bullying and Harassment	Bullying and Harassment: Proactive rate increased from 67% in Q1 2022 to 76.8% in Q2 2022 due to expanding our improved prioritization for proactive detection technology to more languages. Proactive rate increased from 58.8% in Q4 2021 to 67% in Q1 2022 due to the improvement and expansion of our proactive detection technology.

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

		Latest Period						Previous Period						
GARM Category	Relevant Policy	Q2 2022			Q1 2022			Q4 2021			Q3 2021			Commentary
		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	2.5m	465.8k	71.1k	274.3k	47k	239.6k	311.4k	36.6k	261.9k	272.9k	32.9k	233.6k	Adult Nudity and Sexual Activity: Restored content increased from 286.6K in Q1 2022 to 536.8K in Q2 2022 due to the manual restore of a viral, non-violating video. Child Endangerment: Nudity and Physical Abuse: Restored content decreased significantly from 167.9K in Q3 2021 to 19.9K in Q4 2021. This marks a return to pre-Q3 levels following the restore of a large amount of non-violating content in a single-day spike in July. Child Endangerment: Sexual Exploitation: Restored content decreased from 687.8K in Q1 2022 to 17K in Q2 2022, returning to pre-Q4 levels after a period of increased restores for old, non-violating content enforced by our media-matching technology. Restored content increased significantly from 180.5K in Q4 2021 to 687.8K in Q1 2022 after a bug in our media-matching technology was fixed, leading to a jump in restores in March. Restored content increased significantly from 2.8K in Q3 2021 to 180.6K in Q4 2021. In Q4, we reviewed our media-matching technology for old, non-violating content that we restored.
	Child Endangerment: Nudity and Physical Abuse	61.5k	11.1k	18.7k	4k	700	21.2k	3.7k	800	19.2k	2.3k	700	167.2k	
	Child Endangerment: Sexual Exploitation	403k	1.3k	15.7k	800	100	687.8k	800	70	180.5k	700	30	2.8k	
Arms & Ammunition	Regulated Goods: Firearms	149.9k	31.8k	31.7k	74.2k	12.8k	67.5k	63k	10.7k	59.8k	44.1k	9k	60.7k	Regulated Goods: Firearms: Appealed content decreased from 125.3K pieces of content in Q1 2021 to 82.7K in Q2 2021 which is also due to the decline in violating firearms content.
	Violence and Incitement	4.5m	554.5k	6.8k	756k	77.7k	402.8k	361.8k	33.9k	198.8k	435.3k	37.9k	209.5k	

Authorized Metric: Appeals, Reinstatements

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

		Latest Period						Previous Period						
GARM Category	Relevant Policy	Q2 2022			Q1 2022			Q4 2021			Q3 2021			Commentary
		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	2.5m	465.8k	71.1k	274.3k	47k	239.6k	311.4k	36.6k	261.9k	272.9k	32.9k	233.6k	Adult Nudity and Sexual Activity: Restored content increased from 286.6K in Q1 2022 to 536.8K in Q2 2022 due to the manual restore of a viral, non-violating video.
	Violence and Incitement	4.5m	554.5k	6.8k	756k	77.7k	402.8k	361.8k	33.9k	198.8k	435.3k	37.9k	209.5k	Bullying and Harassment: Appealed content decreased from 1 million in Q3 2021 to 799.4K in Q4 2021, following a proportionate decrease in content actioned.
	Violent and Graphic Content	53.8k	8.4k	34.7k	4.5k	800	12k	3.7k	600	5.5k	3.3k	700	8k	Child Endangerment: Nudity and Physical Abuse: Restored content decreased significantly from 167.9K in Q3 2021 to 19.9K in Q4 2021. This marks a return to pre-Q3 levels following the restore of a large amount of non-violating content in a single-day spike in July.
	Bullying and Harassment	1.9m	514.7k	28.7k	736k	114.7k	333.4k	799.4k	121.4k	242.5k	1m	149.8k	282.6k	Child Endangerment: Sexual Exploitation: Restored content decreased from 687.8K in Q1 2022 to 17K in Q2 2022, returning to pre-Q4 levels after a period of increased restores for old, non-violating content enforced by our media-matching technology. Restored content increased significantly from 180.5K in Q4 2021 to 687.8K in Q1 2022 after a bug in our media-matching technology was fixed, leading to a jump in restores in March. Restored content increased significantly from 2.8K in Q3 2021 to 180.6K in Q4 2021. In Q4, we reviewed our media-matching technology for old, non-violating content that we restored.
	Child Endangerment: Nudity and Physical Abuse	61.5k	11.1k	18.7k	4k	700	21.2k	3.7k	800	19.2k	2.3k	700	167.2k	
	Child Endangerment: Sexual Exploitation	403k	1.3k	15.7k	800	100	687.8k	800	70	180.5k	700	30	2.8k	Suicide and Self-Injury: Restored content increased from 345.6K in Q1 2022 to 781.7K in Q2 2022 after we resolved incorrect actions taken by our media-matching technology on non-violating content in June. Restored content increased from 95.3K in Q4 2021 to 345.6K in Q1 2022 after resolving an issue which caused our media-matching technology to action non-violating content. Restored content decreased from 162K in Q3 2021 to 95.3K in Q4 2021, following a period of elevated restores of non-violating, viral content in Q3.
	Suicide and Self-Injury	461k	186.6k	595.1k	6.1k	1.7k	343.9k	200	50	95.2k	200	70	161.5k	Violent and Graphic Content: Restored content increased from 12.8K in Q1 2022 to 43K in Q2 2022 due to the manual restore of incorrectly removed photos of the Russian invasion of Ukraine. Restored content increased from 6.2K in Q4 2021 to 12.8K in Q1 2022 due the automated restore of incorrectly removed photos and videos of the Russian invasion of Ukraine.
	Regulated Goods: Firearms	149.9k	31.8k	31.7k	74.2k	12.8k	67.5k	63k	10.7k	59.8k	44.1k	9k	60.7k	

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

| | | Latest Period | | | | | | Previous Period | | | | | | Commentary |
| | | Q2 2022 | | | Q1 2022 | | | Q4 2021 | | | Q3 2021 | | | |
GARM Category	Relevant Policy	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	53.8k	8.4k	34.7k	4.5k	800	12k	3.7k	600	5.5k	3.3k	700	8k	Violent and Graphic Content: Restored content increased from 12.8K in Q1 2022 to 43K in Q2 2022 due to the manual restore of incorrectly removed photos of the Russian invasion of Ukraine. Restored content increased from 6.2K in Q4 2021 to 12.8K in Q1 2022 due the automated restore of incorrectly removed photos and videos of the Russian invasion of Ukraine. Suicide and Self-Injury: Restored content increased from 345.6K in Q1 2022 to 781.7K in Q2 2022 after we resolved incorrect actions taken by our media-matching technology on non-violating content in June. Restored content increased from 95.3K in Q4 2021 to 345.6K in Q1 2022 after resolving an issue which caused our media-matching technology to action non-violating content. Restored content decreased from 162K in Q3 2021 to 95.3K in Q4 2021, following a period of elevated restores of non-violating, viral content in Q3.
	Violence and Incitement	4.5m	554.5k	6.8k	756k	77.7k	402.8k	361.8k	33.9k	198.8k	435.3k	37.9k	209.5k	
	Suicide and Self Injury	461k	186.6k	595.1k	6.1k	1.7k	343.9k	200	50	95.2k	200	70	161.5k	
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 resinstated 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	2.7m	237.9k	11.8k	586.7k	48.7k	218.2k	768.8k	65.3k	227.8k	1.1m	90.7k	303k	Dangerous Organizations: Organized Hate: Restored content decreased from 231.6K in Q1 2022 to 69.3K in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Restored content increased in Q1 2022 from Q4 2021, due to content that was restored following the bug. Dangerous Organizations: Terrorism: Restored content decreased from 413.7K in Q1 2022 to 86.3K in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Restored content increased in Q1 2022 from Q4 2021, due to content that was restored following the bug. Hate Speech: Appealed content decreased from 769K in Q4 2021 to 587K in Q1 2022, following a proportionate decrease in content actioned. Bullying and Harassment: Appealed content decreased from 1 million in Q3 2021 to 799.4K in Q4 2021, following a proportionate decrease in content actioned.
	Dangerous Organizations: Terrorism	531k	61.9k	24.4k	33.8k	5.4k	408.3k	39k	5.2k	52.3k	64.7k	10.2k	85.2k	
	Dangerous Organizations: Organized Hate	204.9k	60.2k	9.1k	34k	12k	219.6k	35.7k	11.9k	115.4k	53.3k	23.8k	433.1k	
	Bullying and Harassment	1.9m	514.7k	28.7k	736k	114.7k	333.4k	799.4k	121.4k	242.5k	1m	149.8k	282.6k	

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	Q2 2022 Appealed	Q2 2022 Reinstated with appeal	Q2 2022 Reinstated without appeal	Q1 2022 Appealed	Q1 2022 Reinstated with appeal	Q1 2022 Reinstated without appeal	Q4 2021 Appealed	Q4 2021 Reinstated with appeal	Q4 2021 Reinstated without appeal	Q3 2021 Appealed	Q3 2021 Reinstated with appeal	Q3 2021 Reinstated without appeal	Commentary
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	2.7m	237.9k	11.8k	586.7k	48.7k	218.2k	768.8k	65.3k	227.8k	1.1m	90.7k	303k	Hate Speech: Appealed content decreased from 769K in Q4 2021 to 587K in Q1 2022, following a proportionate decrease in content actioned.
	Bullying and Harassment	1.9m	514.7k	28.7k	736k	114.7k	333.4k	799.4k	121.4k	242.5k	1m	149.8k	282.6k	Bullying and Harassment: Appealed content decreased from 1 million in Q3 2021 to 799.4K in Q4 2021, following a proportionate decrease in content actioned.
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	241.5k	44.1k	51.9k	104.1k	37.6k	111.4k	80k	27.7k	119.8k	43.9k	9.3k	83.5k	Regulated Goods: Drugs: Restored content decreased from 149K in Q1 2022 to 96.1K in Q2 2022 due to improvements made to our proactive detection technology. Appealed content increased from 80K in Q4 2021 to 104K in Q1 2022 due to a temporary decrease in the accuracy of enforcement by human reviewers. Appealed content increased from 43.9K in Q3 2021 to 80K in Q4 2021, following a proportionate increase in content actioned.
Spam or Harmful Content	Spam	611.5k	83.8k	117.3m	39.7k	1.7k	33m	21.7k	2k	54.1m	18.7k	1.2k	20.9m	Spam: Restored content increased from 33 million in Q1 2022 to 117 million in Q2 2022 due to a bug in our appeals routing that was later fixed. Restored content decreased from 54.1 million in Q4 2021 to 33 million in Q1 2022, returning to pre-Q4 levels. Restored content increased from 20.9 million in Q3 2021 to 54.1 million in Q4 2021 due to corrections made by our proactive detection technologies.
Terrorism	Dangerous Organizations: Terrorism	531k	61.9k	24.4k	33.8k	5.4k	408.3k	39k	5.2k	52.3k	64.7k	10.2k	85.2k	Dangerous Organizations: Organized Hate: Restored content decreased from 231.6K in Q1 2022 to 69.3K in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Restored content increased in Q1 2022 from Q4 2021, due to content that was restored following the bug.
	Dangerous Organizations: Organized Hate	204.9k	60.2k	9.1k	34k	12k	219.6k	35.7k	11.9k	115.4k	53.3k	23.8k	433.1k	Dangerous Organizations: Terrorism: Restored content decreased from 413.7K in Q1 2022 to 86.3K in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Restored content increased in Q1 2022 from Q4 2021, due to content that was restored following the bug.
Debated Sensitive Social Issue	Hate Speech	2.7m	237.9k	11.8k	586.7k	48.7k	218.2k	768.8k	65.3k	227.8k	1.1m	90.7k	303k	Hate Speech: Appealed content decreased from 769K in Q4 2021 to 587K in Q1 2022, following a proportionate decrease in content actioned.
	Bullying and Harassment	1.9m	514.7k	28.7k	736k	114.7k	333.4k	799.4k	121.4k	242.5k	1m	149.8k	282.6k	Bullying and Harassment: Appealed content decreased from 1 million in Q3 2021 to 799.4K in Q4 2021, following a proportionate decrease in content actioned.

Latest Period: Q2 2022, Q1 2022
Previous Period: Q4 2021, Q3 2021

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 |
		Q2 2022	Q1 2022	Q4 2021	Q3 2021	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	**0.02-0.03%**	**0.02-0.03%**	**0.02-0.03%**	**0.02-0.03%**	Adult Nudity and Sexual Activity: Prevalence remained relatively consistent. 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	
Arms & Ammunition	Regulated Goods: Firearms	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	Violence and Incitement: Prevalence remained relatively consistent.
	Violence and Incitement	**0.01-0.02%**	**0.01-0.02%**	**0.01-0.02%**	**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
		Q2 2022	Q1 2022	Q4 2021	Q3 2021	
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	0.02-0.03%	0.02-0.03%	0.02-0.03%	0.02-0.03%	
	Violence and Incitement	0.01-0.02%	0.01-0.02%	0.01-0.02%	0.02%	
	Violent and Graphic Content	0.01-0.02%	0.01-0.02%	0.01-0.02%	0.01-0.02%	
	Bullying and Harassment	0.04-0.05%	0.05-0.06%	0.05-0.06%	0.05-0.06%	Adult Nudity and Sexual Activity: Prevalence remained relatively consistent.
	Child Endangerment: Nudity and Physical Abuse	N/A	N/A	N/A	N/A	Bullying and Harassment: Prevalence in Q2 2022 decreased due to AI improvements.
	Child Endangerment: Sexual Exploitation	N/A	N/A	N/A	N/A	Violence and Incitement: Prevalence remained relatively consistent.
	Suicide and Self-Injury	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	Violent and Graphic Content: Prevalence remained relatively consistent.
	Regulated Goods: 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
		Q2 2022	Q1 2022	Q4 2021	Q3 2021	
Death, Injury or Military Conflict	Violent and Graphic Content	0.01%-0.02%	0.01%-0.02%	0.01%-0.02%	0.01%-0.02%	Violent and Graphic Content: Prevalence remained relatively consistent.
	Violence and Incitement	0.01-0.02%	0.01-0.02%	0.01-0.02%	0.01-0.02%	Violence and Incitement: Prevalence remained relatively consistent.
	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%	0.02%	0.02-0.03%	0.02%	Hate Speech: Prevalence remained relatively consistent.

Dangerous Organizations: Prevalence remained relatively consistent.

Bullying and Harassment: Prevalence in Q2 2022 decreased due to AI improvements. |
	Dangerous Organizations: Terrorism	Less than 0.05%	Less than 0.05%	Less than 0.06%	Less than 0.05%	
	Dangerous Organizations: Organized Hate	N/A	N/A	N/A	N/A	
	Bullying and Harassment	0.04-0.05%	0.05-0.06%	0.05-0.06%	0.05-0.06%	

GARM Global Alliance for Responsible Media

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.

		Latest Period		Previous Period		
GARM Category	Relevant Policy	Q2 2022	Q1 2022	Q4 2021	Q3 2021	Commentary
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	0.01-0.02%	0.02%	0.02-0.03%	0.02%	Hate Speech: Prevalence remained relatively consistent. Bullying and Harassment: Prevalence in Q2 2022 decreased due to AI improvements.
	Bullying and Harassment	0.04-0.05%	0.05-0.06%	0.05-0.06%	0.05-0.06%	
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: 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	
Terrorism	Dangerous Organizations: Terrorism	Less than 0.05%	Less than 0.05%	Less than 0.06%	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	0.01-0.02%	0.02%	0.02-0.03%	0.02%	Hate Speech: Prevalence remained relatively consistent. Bullying and Harassment: Prevalence in Q2 2022 decreased due to AI improvements.
	Bullying and Harassment	0.04-0.05%	0.05-0.06%	0.05-0.06%	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

<div style="text-align:center">Latest Period   Previous Period</div>

GARM Category	Relevant Policy	Q2 2022		Q1 2022		Q4 2021		Q3 2021	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	10.3m	94.3%	10.4m	94.0%	11.3m	94.3%	10.9m	95.4%
	Child Endangerment: Sexual Exploitation	1.2m	94.9%	1.5m	92.5%	2.6m	97.3%	1.6m	96.3%
	Child Endangerment: Nudity and Physical Abuse	479.8k	93.4%	600.7k	93.8%	983.4k	95.3%	526.5k	92.3%
Arms & Ammunition	Regulated Goods: Firearms	214.8k	93.6%	151k	92.2%	195k	94.3%	154.4k	95.8%
	Violence and Incitement	3.7m	97.0%	2.7m	95.4%	2.6m	96.0%	3.3m	96.4%
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	10.3m	94.3%	10.4m	94.0%	11.3m	94.3%	10.9m	95.4%
	Violence and Incitement	3.7m	97.0%	2.7m	95.4%	2.6m	96.0%	3.3m	96.4%
	Violent and Graphic Content	10.2m	99.3%	6.1m	99.0%	5.5m	98.7%	10.7m	99.3%
	Bullying and Harassment	6.1m	87.4%	7m	83.8%	6.6m	82.1%	7.8m	83.2%
	Child Endangerment: Nudity and Physical Abuse	479.8k	93.4%	600.7k	93.8%	983.4k	95.3%	526.5k	92.3%
	Child Endangerment: Sexual Exploitation	1.2m	94.9%	1.5m	92.5%	2.6m	97.3%	1.6m	96.3%
	Suicide and Self-Injury	6.4m	98.4%	5.1m	98.0%	7.8m	98.4%	3.5m	96.8%
	Regulated Goods: Firearms	214.8k	93.6%	151k	92.2%	195k	94.3%	154.4k	95.8%
Death, Injury or Military Conflict	Violent and Graphic Content	10.2m	99.3%	6.1m	99.0%	5.5m	98.7%	10.7m	99.3%
	Violence and Incitement	3.7m	97.0%	2.7m	95.4%	2.6m	96.0%	3.3m	96.4%
	Suicide and Self Injury	6.4m	98.4%	5.1m	98.0%	7.8m	98.4%	3.5m	96.8%
Online Piracy	Intellectual Property: Copyright	N/A*	N/A*	N/A*	N/A*	N/A*	N/A*	N/A*	N/A*
	Intellectual Property: Counterfeit	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*

* We release the H2 2021 figures in May, and they are not yet available at the time of this report publishing

Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Instagram

		Latest Period						Previous Period	
		Q2 2022		Q1 2022		Q4 2021		Q3 2021	
GARM Category	Relevant Policy	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
Hate speech & acts of aggression	Hate Speech	3.8m	91.2%	3.4m	89.6%	3.8m	91.9%	6m	93.8%
	Dangerous Organizations: Terrorism	1.9m	93.3%	1.5m	86.3%	905.6k	79.6%	685.2k	72.3%
	Dangerous Organizations: Organized Hate	449.1k	87.6%	481.3k	88.9%	332.2k	84.8%	305.8k	82.7%
	Bullying and Harassment	6.1m	87.4%	7m	83.8%	6.6m	82.1%	7.8m	83.2%
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	3.8m	91.2%	3.4m	89.6%	3.8m	91.9%	6m	93.8%
	Bullying and Harassment	6.1m	87.4%	7m	83.8%	6.6m	82.1%	7.8m	83.2%
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	1.9m	96.8%	1.8m	96.0%	1.2m	95.0%	1.8m	97.4%
Spam or Harmful Content	Spam	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Terrorism	Dangerous Organizations: Terrorism	1.9m	93.3%	1.5m	86.3%	905.6k	79.6%	685.2k	72.3%
	Dangerous Organizations: Organized Hate	449.1k	87.6%	481.3k	88.9%	332.2k	84.8%	305.8k	82.7%
Debated Sensitive Social Issue	Hate Speech	3.8m	91.2%	3.4m	89.6%	3.8m	91.9%	6m	93.8%
	Bullying and Harassment	6.1m	87.4%	7m	83.8%	6.6m	82.1%	7.8m	83.2%

GARM Global Alliance for Responsible Media

Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Instagram

GARM Category	Relevant Policy	Commentary
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	
	Child Endangerment: Sexual Exploitation	Proactive rate increased from 92.5% in Q1 2022 to 94.9% in Q2 2022, returning to pre-Q1 levels after a spike in reported viral links in February. Content actioned decreased from 2.6 million in Q4 2021 to 1.5 million in Q1 2022, returning to pre-Q4 levels following an update to our media-matching technology that had enabled us to take down more old, violating content. Proactive rate decreased from 97.3% in Q4 2021 to 92.5% in Q1 2022 due to a spike in reported viral links in February. Content actioned increased from 1.6 million in Q3 2021 to 2.6 million in Q4 2021 as we used our media-matching technology to identify old, violating content.
	Child Endangerment: Nudity and Physical Abuse	Content actioned decreased from 600.8K in Q1 2022 to 479.8K in Q2 2022 due to a change in the operational guidelines that led to actioned content being misclassified. This was resolved in June. Content actioned decreased from 983.4K in Q4 2021 to 600.8K in Q1 2022, returning to pre-Q4 levels following enforcement on old, violating content in Q1 from improvements we made to our proactive detection technology on videos. Content actioned increased significantly from 526.5K in Q3 2021 to 983.4K in Q4 2021. This is because we improved our proactive detection technology on videos, improving the accuracy of our enforcement actions. Proactive rate increased from 92.3% in Q3 2021 to 95.3% in Q4 2021. This is because we improved our proactive detection technology on videos improving the accuracy of our enforcement actions.
Arms & Ammunition	Regulated Goods: Firearms	Content actoned increased from 151K in Q1 2022 to 214.8K in Q2 2022 due to improvements made to our proactive detection technology. Content actioned increased from 76.3K in Q2 2021 to 154.4K in Q3 2021 due to improved and expanded proactive detection technologies. Proactive rate increased from 90.6% in Q2 2021 to 958% in Q3 2021 due to improved and expanded proactive detection technologies. Content actioned increased from 154.4K in Q3 2021 to 195K in Q4 2021 due to improved and expanded proactive detection technologies.
	Violence and Incitement	Content actioned increased from 2.7 million in Q1 2022 to 3.7 million in Q2 2022 due to the improvement and expansion of our proactive detection technology.
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	
	Violence and Incitement	Content actioned increased from 2.7 million in Q1 2022 to 3.7 million in Q2 2022 due to the improvement and expansion of our proactive detection technology.
	Violent and Graphic Content	Content actioned increased from 6.1 million in Q1 2022 to 10.2 million in Q2 2022 due to a bug in our proactive detection technology for photos which misclassified and added warning screens to non-violating content. Content actioned decreased from 10.7 million in Q3 2021 to 5.5 million in Q4 2021 following a period of elevated enforcement in Q3 on viral, violating content.
	Bullying and Harassment	Bullying and Harassment: Proactive rate increased from 83.8% in Q1 2022 to 87.4% in Q2 2022 due to expanding our improved prioritization for proactive detection technology to more languages. Content actioned increased from 4.5 million to 7.8 million as we expanded our proactive detection technology to more languages.
	Child Endangerment: Nudity and Physical Abuse	Content actioned decreased from 600.8K in Q1 2022 to 479.8K in Q2 2022 due to a change in the operational guidelines that led to actioned content being misclassified. This was resolved in June. Content actioned decreased from 983.4K in Q4 2021 to 600.8K in Q1 2022, returning to pre-Q4 levels following enforcement on old, violating content in Q1 from improvements we made to our proactive detection technology on videos. Content actioned increased significantly from 526.5K in Q3 2021 to 983.4K in Q4 2021. This is because we improved our proactive detection technology on videos, improving the accuracy of our enforcement actions. Proactive rate increased from 92.3% in Q3 2021 to 95.3% in Q4 2021. This is because we improved our proactive detection technology on videos improving the accuracy of our enforcement actions.
	Child Endangerment: Sexual Exploitation	Proactive rate increased from 92.5% in Q1 2022 to 94.9% in Q2 2022, returning to pre-Q1 levels after a spike in reported viral links in February. Content actioned decreased from 2.6 million in Q4 2021 to 1.5 million in Q1 2022, returning to pre-Q4 levels following an update to our media-matching technology that had enabled us to take down more old, violating content. Proactive rate decreased from 97.3% in Q4 2021 to 92.5% in Q1 2022 due to a spike in reported viral links in February. Content actioned increased from 1.6 million in Q3 2021 to 2.6 million in Q4 2021 as we used our media-matching technology to identify old, violating content.
	Suicide and Self-Injury	Content actioned increased from 5.1 million in Q1 2022 to 6.4 million in Q2 2022 due to an increase in automated enforcement on non-violating content added to our media-matching technology banks incorrectly by human reviewers. Content actioned decreased from 7.8 million in Q4 2021 to 5.1 million in Q1 2022 following updates made to our media-matching technology in Q4 which led to the removal of old violating content.
	Regulated Goods: Firearms	Content actioned increased from 76.3K in Q2 2021 to 154.4K in Q3 2021 due to improved and expanded proactive detection technologies. Proactive rate increased from 90.6% in Q2 2021 to 958% in Q3 2021 due to improved and expanded proactive detection technologies.
Death, Injury or Military Conflict	Violent and Graphic Content	Content actioned increased from 6.1 million in Q1 2022 to 10.2 million in Q2 2022 due to a bug in our proactive detection technology for photos which misclassified and added warning screens to non-violating content. Content actioned decreased from 10.7 million in Q3 2021 to 5.5 million in Q4 2021 following a period of elevated enforcement in Q3 on viral, violating content.
	Violence and Incitement	Content actioned increased from 2.7 million in Q1 2022 to 3.7 million in Q2 2022 due to the improvement and expansion of our proactive detection technology.
	Suicide and Self Injury	Content actioned increased from 5.1 million in Q1 2022 to 6.4 million in Q2 2022 due to an increase in automated enforcement on non-violating content added to our media-matching technology banks incorrectly by human reviewers. Content actioned decreased from 7.8 million in Q4 2021 to 5.1 million in Q1 2022 following updates made to our media-matching technology in Q4 which led to the removal of old violating content.

GARM Global Alliance for Responsible Media

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
Online piracy	Intellectual Property: Copyright	We report this metric monthly in a 6 month report. Our current report has data for January - June 2021. 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. Our proactive rate figure here constitutes the volume of content removed in response to an IP report relative to the volume of content reported, reflected as a percentage. In prior transparency reports, the Removal Rate constituted the percentage of total IP reports that resulted in some or all reported content being removed. Beginning in the July 2019 reporting period, we have adjusted the way we calculate Removal Rate to reflect the percentage of reported content removed, rather than the percentage of reports resulting in removals. Because a single IP report can identify multiple pieces of content, this figure offers a more complete picture of the total content removed from the platform based on an IP report.
	Intellectual Property: Counterfeit	
	Intellectual Property: Trademark	
Hate speech & acts of aggression	Hate Speech	Content actioned for hate speech decreased from 9.8 million in Q2 2021 to 6 million in Q3 2021 back to pre-Q2 levels.
	Dangerous Organizations: Terrorism	Content actioned increased from 1.5 million in Q1 2022 to 1.9 million in Q2 2022 primarily due to a jump in violating content from India actioned by our media-matching technology in June. Proactive rate increased from 86.3% in Q1 2022 to 93.3% in Q2 2022 due to an increase in proactive actions taken by our media-matching technology. Content actioned increased from 336.9K pieces of content in Q2 2021 to 398.8K in Q3 2021 due to adjustments to our proactive detection technologies. Content actioned increased from 685.2K pieces of content in Q3 2021 to 905.3K in Q4 2021 due to improvements made to our proactive detection technologies.
	Dangerous Organizations: Organized Hate	Content actioned increased in Q1 2022 from Q4 2021 for both Terrorism and Organized Hate due to an increase in enforcement on non-violating content due to a bug in our media-matching technology that was later fixed and the content was restored. Content actioned increased from 306K in Q3 2021 to 332K in Q4 2021 due to updates made to our proactive detection technology we launched in early Q4 2021. Proactive rate increased from 82.7% in Q3 2021 to 84.8% in Q4 2021 due to improvements made to our media-matching technology which allowed us to detect and remove old, violating content.
	Bullying and Harassment	Bullying and Harassment: Proactive rate increased from 83.8% in Q1 2022 to 87.4% in Q2 2022 due to expanding our improved prioritization for proactive detection technology to more languages. Content actioned increased from 4.5 million to 7.8 million as we expanded our proactive detection technology to more languages.
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	Content actioned for hate speech decreased from 9.8 million in Q2 2021 to 6 million in Q3 2021 back to pre-Q2 levels.
	Bullying and Harassment	Bullying and Harassment: Proactive rate increased from 83.8% in Q1 2022 to 87.4% in Q2 2022 due to expanding our improved prioritization for proactive detection technology to more languages. Content actioned increased from 4.5 million to 7.8 million as we expanded our proactive detection technology to more languages.
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	Content actioned increased from 1.2 million in Q4 2021 to 1.8 million in Q1 2022 due to updates made to our proactive detection technologies. Content actioned decreased from from 1.8 million in Q3 2021 to 1.2 million in Q4 2021 due to updates made to proactive detection technologies.
Spam or Harmful Content	Spam	
Terrorism	Dangerous Organizations: Terrorism	Content actioned increased from 1.5 million in Q1 2022 to 1.9 million in Q2 2022 primarily due to a jump in violating content from India actioned by our media-matching technology in June. Proactive rate increased from 86.3% in Q1 2022 to 93.3% in Q2 2022 due to an increase in proactive actions taken by our media-matching technology. Content actioned increased from 336.9K pieces of content in Q2 2021 to 398.8K in Q3 2021 due to adjustments to our proactive detection technologies. Content actioned increased from 685.2K pieces of content in Q3 2021 to 905.3K in Q4 2021 due to improvements made to our proactive detection technologies.
	Dangerous Organizations: Organized Hate	Content actioned increased in Q1 2022 from Q4 2021 for both Terrorism and Organized Hate due to an increase in enforcement on non-violating content due to a bug in our media-matching technology that was later fixed and the content was restored. Content actioned increased from 306K in Q3 2021 to 332K in Q4 2021 due to updates made to our proactive detection technology we launched in early Q4 2021. Proactive rate increased from 82.7% in Q3 2021 to 84.8% in Q4 2021 due to improvements made to our media-matching technology which allowed us to detect and remove old, violating content.
Debated Sensitive Social Issue	Hate Speech	Content actioned for hate speech decreased from 9.8 million in Q2 2021 to 6 million in Q3 2021 back to pre-Q2 levels.
	Bullying and Harassment	Bullying and Harassment: Proactive rate increased from 83.8% in Q1 2022 to 87.4% in Q2 2022 due to expanding our improved prioritization for proactive detection technology to more languages. Content actioned increased from 4.5 million to 7.8 million as we expanded our proactive detection technology to more languages.

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

| | | Latest Period | | | | | | Previous Period | | | | | | |
| | | Q2 2022 | | | Q1 2022 | | | Q4 2021 | | | Q3 2021 | | | |
GARM Category	Relevant Policy	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	Commentary
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	867.4k	185.8k	80.3k	0	0	183.8k	0	0	227.8k	0	0	159.2k	Adult Nudity and Sexual Activity: Restored content increased from 183.8K in Q1 2022 to 266K in Q2 2022 due to a bug in our proactive detection technology that was later fixed. Restored content increased from 159.2K in Q3 2021 to 227.8K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.\n\nChild Endangerment: Sexual Exploitation: Restored content decreased from 154.4K in Q1 2022 to 613 in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology was fixed, leading to a jump in restores in March.
	Child Endangerment: Sexual Exploitation	3.9k	200	400	0	20	154.2k	0	0	1.6k	0	0	300	
	Child Endangerment: Nudity and Physical Abuse	29.1k	3.8k	5.8k	0	0	10.7k	0	0	13.6k	0	0	168.3k	
Arms & Ammunition	Regulated Goods: Firearms	21.6k	13.8k	6.2k	0	0	15.4k	0	0	21.7k	0	0	7.8k	Violence and Incitement: Restored content increased from 21.4K in Q3 2021 to 49.9K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.
	Violence and Incitement	327k	54.3k	9.9k	0	0	53.1k	0	0	49.9k	0	0	21.4k	

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

		Latest Period						Previous Period						
GARM Category	Relevant Policy	Q2 2022			Q1 2022			Q4 2021			Q3 2021			Commentary
		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	867.4k	185.8k	80.3k	0	0	183.8k	0	0	227.8k	0	0	159.2k	Adult Nudity and Sexual Activity: Restored content increased from 183.8K in Q1 2022 to 266K in Q2 2022 due to a bug in our proactive detection technology that was later fixed. Restored content increased from 159.2K in Q3 2021 to 227.8K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.

Violence and Incitement: Restored content increased from 21.4K in Q3 2021 to 49.9K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.

Violent and Graphic Content: Restored content increased from 31.5K pieces of content in Q1 2022 to 1.1 million in Q2 2022 after a bug in our proactive detection technology was fixed, leading to an increase in restores of non-violating content. Restored content decreased from 2.5 million pieces of content in Q4 2021 to 31.5K in Q1 2022, returning to pre-Q4 levels following the automated restore of a viral, non-violating image. Restored content increased from 21.2K pieces of content in Q3 2021 to 2.5 million in Q4 2021 due to the automated restore of a viral, non-violating image.

Bullying and Harassment: Restored content increased from 91.5K in Q3 2021 to 182.1K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.

Child Nudity and Sexual Exploitation: From Q2 2021 we created two new reporting categories under the broader topic of child endangerment; Child Endangerment: Nudity and Physical Abuse and Child Endangerment: Sexual Exploitation.

Child Endangerment: Nudity and Physical Abuse -- Restored content for child nudity and physical abuse increased significantly from 4.5K in Q2 2021 to 168.2K in Q3 2021. This is because we restored a large amount of non-violating content in late July.

Suicide and Self-Injury: Restored content increased from 49.4K in Q1 2022 to 118.8K in Q2 2022 after we resolved incorrect actions taken by our media-matching technology on non-violating content in June. Restored content increased from 624 in Q4 2021 to 49.4K in Q1 2022 after resolving a technical issue which prevented users from being able to use the "disagree with decision" option when content was actioned. Restored content decreased from 10.6K in Q3 2021 to 624 in Q4 2021, following a period of elevated restores of viral, non-violating content in Q3.

Regulated Goods: Firearms Restored content increased from 15.4K in Q1 2022 to 20K in Q2 2022, following a proportionate increase in content actioned. |
	Violence and Incitement	327k	54.3k	9.9k	0	0	53.1k	0	0	49.9k	0	0	21.4k	
	Violent and Graphic Content	157.9k	19.7k	1.1m	0	0	31.4k	0	0	2.5m	0	0	21.2k	
	Bullying and Harassment	853.5k	168.6k	18.1k	0	0	215.8k	0	0	182.1k	0	0	91.5k	
	Child Endangerment: Nudity and Physical Abuse	29.1k	3.8k	5.8k	0	0	10.7k	0	0	13.6k	0	0	168.3k	
	Child Endangerment: Sexual Exploitation	3.9k	200	400	0	20	154.2k	0	0	1.6k	0	0	300	
	Suicide and Self-Injury	177.6k	79.2k	39.8k	0	0	49.4k	0	0	600	0	0	10.6k	
	Regulated Goods: Firearms	21.6k	13.8k	6.2k	0	0	15.4k	0	0	21.7k	0	0	7.8k	

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

		Latest Period						Previous Period						Commentary
		Q2 2022			Q1 2022			Q4 2021			Q3 2021			
GARM Category	Relevant Policy	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	157.9k	19.7k	1.1m	0	0	31.4k	0	0	2.5m	0	0	21.2k	Violent and Graphic Content: Restored content increased from 31.5K pieces of content in Q1 2022 to 1.1 million in Q2 2022 after a bug in our proactive detection technology was fixed, leading to an increase in restores of non-violating content. Restored content decreased from 2.5 million pieces of content in Q4 2021 to 31.5K in Q1 2022, returning to pre-Q4 levels following the automated restore of a viral, non-violating image. Restored content increased from 21.2K pieces of content in Q3 2021 to 2.5 million in Q4 2021 due to the automated restore of a viral, non-violating image. Violence and Incitement: Restored content increased from 21.4K in Q3 2021 to 49.9K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.
	Violence and Incitement	327k	54.3k	9.9k	0	0	53.1k	0	0	49.9k	0	0	21.4k	
	Suicide and Self Injury	177.6k	79.2k	39.8k	0	0	49.4k	0	0	600	0	0	10.6k	Suicide and Self-Injury: Restored content increased from 49.4K in Q1 2022 to 118.8K in Q2 2022 after we resolved incorrect actions taken by our media-matching technology on non-violating content in June. Restored content increased from 624 in Q4 2021 to 49.4K in Q1 2022 after resolving a technical issue which prevented users from being able to use the "disagree with decision" option when content was actioned. Restored content decreased from 10.6K in Q3 2021 to 624 in Q4 2021, following a period of elevated restores of viral, non-violating content in Q3.
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	
	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	We do not report content appealed and resinstated of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	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	397.4k	47.8k	14.1k	0	0	56.7k	0	0	63.6k	0	0	43.1k	Hate Speech: Restored content increased from 43.1K in Q3 2021 to 63.6K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool. Dangerous Organizations: Organized Hate: Restored content decreased from 31.2K in Q1 2022 to 20.9K in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Restored content increased in Q1 2022 from Q4 2021, due to content that was restored following the bug. Dangerous Organizations: Terrorism: Restored content decreased from 84.9K in Q1 2022 to 22.5K in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Restored content increased in Q1 2022 from Q4 2021, due to content that was restored following the bug. Bullying and Harassment: Restored content increased from 91.5K in Q3 2021 to 182.1K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.
	Dangerous Organizations: Terrorism	99.4k	18k	4.5k	0	0	84.9k	0	0	4k	0	0	3.5k	
	Dangerous Organizations: Organized Hate	46.9k	14.7k	6.2k	0	0	31.2k	0	0	7.5k	0	0	3.7k	
	Bullying and Harassment	853.5k	168.6k	18.1k	0	0	215.8k	0	0	182.1k	0	0	91.5k	

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
		Q2 2022			Q1 2022			Q4 2021			Q3 2021			
		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	397.4k	47.8k	14.1k	0	0	56.7k	0	0	63.6k	0	0	43.1k	Hate Speech: Restored content increased from 43.1K in Q3 2021 to 63.6K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.
	Bullying and Harassment	853.5k	168.6k	18.1k	0	0	215.8k	0	0	182.1k	0	0	91.5k	Bullying and Harassment: Restored content increased from 91.5K in Q3 2021 to 182.1K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	160k	22.4k	5.2k	0	0	45k	0	0	27.6k	0	0	35.6k	Restored content decreased from 45K in Q1 2022 to 27.6K in Q2 2022 due to improvements made to our proactive detection technology. Restored content increased from 27.6K in Q4 2021 to 45K in Q1 2022, following a proportionate increase in content actioned.
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	99.4k	18k	4.5k	0	0	84.9k	0	0	4k	0	0	3.5k	Dangerous Organizations: Organized Hate: Restored content decreased from 31.2K in Q1 2022 to 20.9K in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Restored content increased in Q1 2022 from Q4 2021, due to content that was restored following the bug.
	Dangerous Organizations: Organized Hate	46.9k	14.7k	6.2k	0	0	31.2k	0	0	7.5k	0	0	3.7k	Dangerous Organizations: Terrorism: Restored content decreased from 84.9K in Q1 2022 to 22.5K in Q2 2022 returning to pre-Q1 levels after a bug in our media-matching technology in Q1 was fixed. Restored content increased in Q1 2022 from Q4 2021, due to content that was restored following the bug.
Debated Sensitive Social Issue	Hate Speech	397.4k	47.8k	14.1k	0	0	56.7k	0	0	63.6k	0	0	43.1k	Hate Speech: Restored content increased from 43.1K in Q3 2021 to 63.6K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.
	Bullying and Harassment	853.5k	168.6k	18.1k	0	0	215.8k	0	0	182.1k	0	0	91.5k	Bullying and Harassment: Restored content increased from 91.5K in Q3 2021 to 182.1K in Q4 2021 as we provided more transparency into removal decisions in the Account Status tool.

Twitter

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.

Since 2012, the Twitter Transparency Report has offered a window into our work to enforce the Twitter Rules, protect privacy, navigate increased requests from governments around the world, disrupt state-backed information operations, and more. Now, amid an increasingly complex global landscape, continued transparency around our own efforts to protect the public conversation is paramount.

Our latest Twitter Transparency Center update includes data from July 1st to December 31st, 2021. While we continue to share data across consistent, recurring categories, we're introducing new data that can provide meaningful insights into the impact of our actions.

In this Transparency Report, we added a new metric quantifying removals under our Manipulated Media policy. Our Manipulated Media policy is just one part of our efforts to make sure people can engage in the public conversation safely and with confidence, by restricting synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm. In this reporting period, 25 accounts were actioned and 96 pieces of content were removed for violating our Manipulated Media policy. You can learn more about Twitter's approach to addressing misleading content here.

Impressions

We continue to share Impressions data on violative Tweets, a metric we first introduced in 2020. Our Impressions metric captures the number of views a violative Tweet receives prior to removal.

Of the Tweets removed during July 1st, 2021 through December 31st 2021, 71% received fewer than 100 impressions prior to removal, with an additional 21% receiving between 100 and 1,000 impressions. Only 8% of removed Tweets had more than 1,000 impressions. In total, impressions on these violative Tweets accounted for less than 0.1% of all impressions for all Tweets during this reporting period.

We continue to invest into improving these numbers over time, taking enforcement action on violative content before it's even viewed.

We continue to uplevel our proactive enforcement across the service and invest in technological solutions to respond to ever-evolving malicious online activity. In H2 2021, we used technology to proactively identify 92% of accounts that were actioned for violating Twitter's policies on terrorism and violent extremism.

Additionally, we suspended 596,997 unique accounts for violating Twitter's zero-tolerance policy on CSE content during this reporting period – a 32% increase since our previous report. Consistent with previous reporting periods, 91% of these suspended accounts were identified proactively by employing internal proprietary tools and industry hash sharing initiatives, such as PhotoDNA.

Some other notable changes since our last report are an 84% decrease in the number of accounts actioned for violations of our Civic Integrity policy, and a 14% decrease in the number of accounts actioned for violations of our COVID-19 misleading information policy during this reporting period. These changes coincide with a decrease in conversations related to COVID-19 on the platform, as well as the absence of a US election cycle. Additionally, we saw a 10% decrease from our last report in accounts suspended for violating our Abusive Behavior policy. This is in line with an 11% decrease in accounts reported under this policy during this period.

Partnerships

We believe that a safer Twitter is a better Twitter and we're committed to transparently working towards making Twitter a safer place for everyone - both people and brands. As part of that mission, we remain committed to completing the Media Ratings Council's accreditation process across all four of their Accreditation Service focus areas: Viewability, Sophisticated Invalid Traffic Filtration, Audience Measurement, and Brand Safety. In our initial announcement of that commitment, we shared that we would begin this process with the Brand Safety audit, and we continue to engage with the MRC to fulfill that commitment – at the time of writing, we have completed the MRC pre-assessment, and are looking forward to sharing findings with the MRC Audit Committee and progressing to the formal audit.

Additionally, we successfully earned the Trustworthy Accountability Group (TAG) Brand Safety Certified Seal during the first half of 2021 (our previous reporting period), and are engaged with TAG to renew this certification for the second time.

Lastly, we're proud of the progress made in partnership with GARM on the incorporation of misinformation as a twelfth content category in GARM's framework this year. This is a critically important area on which we've provided enforcement transparency for several years, and this partnership has allowed us to make meaningful progress on this front towards greater industry-wide visibility and transparency.

We're committed to increasing our transparency and improving our accountability to the public, and we'll continue to publish updates to the Twitter Transparency Center on a biannual basis.

October 2022

Next Best Measure: Impressions

Number of views a violative Tweet received prior to removal by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H2 2021	Previous Period – H1 2021	Commentary
		Impressions	Impressions	
Adult & explicit sexual content	Non-consensual nudity			
	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	In total, impressions on these violative Tweets accounted for less than 0.1% of all impressions for all Tweets during that time period.	From January 1, 2021 through June 30, 2021, Twitter removed 4.7M Tweets that violated the Twitter Rules. Of the Tweets removed, 68% received fewer than 100 impressions prior to removal, with an additional 24% receiving between 100 and 1,000 impressions. Only 8% of removed Tweets had more than 1,000 impressions. In total, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets globally, from January 1, 2021 through June 30, 2021.	From July 1, 2021 through December 31, 2021, Twitter required users to remove 4M Tweets that violated the Twitter Rules. Of the Tweets removed, 71% received fewer than 100 impressions prior to removal, with an additional 21% receiving between 100 and 1,000 impressions. Only 8% of removed Tweets had more than 1,000 impressions.
	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			

Next best measure: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H2 2021			Previous Period – H1 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Adult & explicit sexual content	Non-consensual nudity	28,836	8,141	60,816	29,635	7,519	64,596	We continue to uplevel our proactive enforcement across the service and invest in technological solutions to respond to ever-evolving malicious online activity. In H2 2021, we used technology to proactively identify 92% of accounts that were actioned for violating Twitter's policies on terrorism and violent extremism. Additionally, we suspended 596,997 unique accounts for violating Twitter's zero-tolerance policy on CSE content during this reporting period – a 32% increase since our previous report. Consistent with previous reporting periods, 91% of these suspended accounts were identified proactively by employing internal proprietary tools and industry hash sharing initiatives, such as PhotoDNA.

Some other notable changes since our last report are an 84% decrease in the number of accounts actioned for violations of our Civic Integrity policy, and a 14% decrease in the number of accounts actioned for violations of our COVID-19 misleading information policy during this reporting period. These changes coincide with a decrease in conversations related to COVID-19 on the platform, as well as the absence of a US election cycle. Additionally, we saw a 10% decrease from our last report in accounts suspended for violating our Abusive Behavior policy. This is in line with an 11% decrease in accounts reported under this policy during this period. |
	Sensitive media	1,143,064	118,356	1,149,829	1,630,554	164,260	1,655,608	
	Child sexual exploitation	599,523	596,997	6,796	456,146	453,754	6,087	
Arms & ammunition	Illegal or certain regulated goods or services	224,185	119,508	571,902	175,798	87,530	420,950	
Crime & harmful acts to individuals and society, human right violations	Violence	61,358	41,386	70,229	89,245	66,445	101,907	
	Abuse/harassment	940,679	82,971	1,344,061	1,043,525	99,565	1,547,654	
Death, injury or military conflict	Promoting suicide or self-harm	408,143	10,197	509,776	345,100	8,621	413,769	
Online piracy	Copyright	Notices Issued: 146,906	Accounts Affected: 623,576	Tweets Withheld: 161,983	Notices Issued: 171,747	Accounts Affected: 796,506	Tweets Withheld: 432,759	
	Trademark	Trademark Notices: 26,274			Trademark Notices: 20,121			
Hate speech & acts of aggression	Hateful conduct	902,169	104,565	1,293,178	1,108,722	133,585	1,606,979	
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	1,143,064	118,356	1,149,829	1,630,554	164,260	1,655,608	
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services	224,185	119,508	571,902	175,798	87,530	420,950	
Spam or harmful content	Private information	34,181	2,563	62,537	30,714	3,178	54,590	
	Impersonation	181,644	169,396	15,275	216,846	199,229	21,188	
	Platform manipulation	Anti-Spam Challenges Issued: 133,266,534			Anti-Spam Challenges Issued: 130,289,899			
Terrorism	Terrorism/violent extremism	33,694	33,693	1	44,974	44,974		
Debated sensitive social issues	N/A							
Other	Civic integrity	93	4	102	581	23	593	
	COVID-19 misleading information	24,012	1,376	30,190	27,935	617	33,761	

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 – H2 2021			Previous Period – H1 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Adult & explicit sexual content	Non-consensual nudity	28,836	8,141	60,816	29,635	7,519	64,596	There was a 3% decrease in the number of accounts actioned for violations of our non-consensual nudity policy during this reporting period.
	Sensitive media	1,143,064	118,356	1,149,829	1,630,554	164,260	1,655,608	There was a 30% decrease in the number of accounts actioned for violations of our sensitive media policy during this reporting period. We removed a total of 1.1M unique pieces of content under our Sensitive Media policy during this period, a 31% decrease since our last report.
	Child sexual exploitation	599,523	596,997	6,796	456,146	453,754	6,087	There was a 31% 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 Twitter Rules regarding Child Sexual Exploitation via our web form.
Arms & ammunition	Illegal or certain regulated goods or services	224,185	119,508	571,902	175,798	87,530	420,950	There was a 28% increase in the number of accounts actioned for violations of our illegal or certain regulated goods or services policy during this reporting period. Due to continued refinement of enforcement guidelines, we saw a 37% increase in accounts suspended under this policy, representing a total of 119,508 accounts.
Crime & harmful acts to individuals and society, human right violations	Violence	61,358	41,386	70,229	89,245	66,445	101,907	There was a 31% 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. 41,386 accounts were suspended and we took action on 70,229 unique pieces of content during this reporting period.
	Abuse/harassment	940,679	82,971	1,344,061	1,043,525	99,565	1,547,654	There was a 10% decrease 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 940,679 accounts during this reporting period. This is a 10% decrease from our last report and is in line with a 11% decrease in accounts reported under this policy during this period.

Global Alliance for Responsible Media

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 – H2 2021			Previous Period – H1 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Death, injury or military conflict	Promoting suicide or self-harm	408,143	10,197	509,776	345,100	8,621	413,769	There was an 18% increase in the number of accounts actioned for violations of our suicide or self-harm policy during this reporting period. We prohibit content that promotes, or otherwise encourages, suicide or self-harm. During this reporting period there was a substantial increase in the volume of accounts suspended (18%), and content removed (23%) under this policy. 408,143 accounts were actioned in total. We attribute these changes to our continued investment in identifying violative content at scale.
Online piracy	Copyright	Notices Issued: 146,906	Accounts Affected: 623,576	Tweets Withheld: 161,983	Notices Issued: 171,747	Accounts Affected: 796,506	Tweets Withheld: 432,759	We saw a 16% decrease in DMCA takedown notices submitted, and a 22% decrease in accounts affected. Tweets withheld dropped by 63% while media withheld decreased by 18%.
	Trademark	Trademark Notices: 26,274			Trademark Notices: 20,121			Twitter received 31% more trademark notices, affecting 8% more accounts since our last report. We carefully review each report received under our trademark policy, 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	902,169	104,565	1,293,178	1,108,722	133,585	1,606,979	There was a 19% decrease in the number of accounts actioned for violations of our hateful conduct policy during this reporting period. We expanded our Hateful Conduct policy in December 2021 to prohibit dehumanizing speech on the basis of gender, gender identity and sexual orientation. During this period 104,565 accounts were suspended under this policy, representing a 22% decrease in account suspensions since our last report.
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	1,143,064	118,356	1,149,829	1,630,554	164,260	1,655,608	There was a 30% decrease in the number of accounts actioned for violations of our sensitive media policy during this reporting period. We removed a total of 1.1M unique pieces of content under our Sensitive Media policy during this period, a 31% decrease since our last report.

GARM Global Alliance for Responsible Media

53

Authorized Metric: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H2 2021			Previous Period – H1 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services	224,185	119,508	571,902	175,798	87,530	420,950	There was a 28% increase in the number of accounts actioned for violations of our illegal or certain regulated goods or services policy during this reporting period. Due to continued refinement of enforcement guidelines, we saw a 37% increase in accounts suspended under this policy, representing a total of 119,508 accounts.
Spam or harmful content	Private information	34,181	2,563	62,537	30,714	3,178	54,590	There was an 11% increase in the number of accounts actioned for violations of our private information policy during this reporting period. We expanded our private information policy in late November to prohibit sharing media of private individuals without the permission of those depicted. 34,181 accounts and 62,537 unique pieces of content were actioned under this policy.
	Impersonation	181,644	169,396	15,275	216,846	199,229	21,188	There was a 16% decrease in the number of accounts actioned for violations of our impersonation policy during this time period. This reporting period, we actioned 181,644 accounts and suspended 169,396 accounts, a 16% and 15% decrease respectively, for violations of the impersonation policy. This decrease is in line with a similar 15% decrease in accounts reported during this period.
	Platform manipulation	Anti-Spam Challenges Issued: **133,266,534**			Anti-Spam Challenges Issued: **130,289,899**			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. These anti-spam challenges increased by approximately 2% compared to the previous reporting period. This nominal increase is related to ongoing efforts to disrupt spam attacks on our platform. During the second half of 2021, we observed an approximately 6% increase in the number of spam reports from the previous reporting period. World events can cause spam reports to fluctuate as users may block and report one another during conversations, and we believe that this increase may be largely correlated with various socio-political events that took place during this time.

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 – H2 2021			Previous Period – H1 2021			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Terrorism	Terrorism/violent extremism	33,694	33,693	1	44,974	44,974		There was a 25% decrease in the number of accounts actioned for violations of our terrorism/violent extremism policy during this reporting period. We suspended 33,693 unique accounts for violations of the policy during this reporting period. Of those accounts, 92% were proactively identified and actioned. Our current methods of surfacing potentially violating content for review include leveraging the shared industry hash database supported by the Global Internet Forum to Counter Terrorism (GIFCT).
Debated sensitive social issues	N/A							
Other	Civic integrity	93	4	102	581	23	593	There was an 84% decrease in the number of accounts actioned for violations of our civic integrity policy during this reporting period. During this reporting period the number of accounts actioned under Civic Integrity policy has decreased due to the low number of major national elections in the United States.
	COVID-19 misleading information	24,012	1,376	30,190	27,935	617	33,761	There was a 14% decrease in the number of accounts actioned for violations or our COVID-19 misleading information policy during this reporting period. This number does not include accounts where we applied a label or warning message. As of March 2021, we incorporated a five-strike system meant to address repeated violations of the COVID-19 misinformation policy. After the fifth strike, the user is eligible for suspension under the policy. Since the launch of the strike system we invested in and increased our proactive detection efforts to surface and mitigate the harm related to COVID-19 misinformation. We suspended 1,376 accounts, an increase of 123%, for violations of the COVID-19 misinformation policy during this reporting period.

Next Best Measure: Proactive Action Rate

Violating content proactively detected by Twitter (without reliance on user reports)

GARM Category	Relevant Twitter Policy	Proactive Action Rate
Adult & explicit sexual content	Non-consensual nudity	
	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	We continue to step up the level of proactive enforcement across the service and invest in technological solutions to respond to ever-evolving malicious online activity, and report proactive enforcement rates at our discretion.
	Trademark	
Hate speech & acts of aggression	Hateful conduct	In H2 2021, we used technology to proactively identify 92% of accounts that were actioned for violating Twitter's policies on terrorism and violent extremism. Additionally, we suspended 596,997 unique accounts for violating Twitter's zero-tolerance policy on CSE content during this reporting period – a 32% increase since our previous report. 91% of these suspended accounts were identified proactively by employing internal proprietary tools and industry hash sharing initiatives, such as PhotoDNA.
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	This is a metric that we report at our discretion, and we may not disclose this metric every reporting period.
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	

GARM Global Alliance for Responsible Media

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	
	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	Twitter does not report appeals data at this time.
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	

TikTok

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. Our Community Guidelines Enforcement Reports provide quarterly insights into the volume and nature of content and accounts removed from our platform.

Our Continued Progress

With each Community Guidelines Enforcement Report, we have continued to expand our reporting to bring ever-more transparency to our actions, progress, and challenges, and to stay accountable to our community. For instance, since the start of 2022, we began to report on the 30 markets with the largest volumes of removed videos, which account for approximately 80% of overall video removal volume. There are also new charts for spam account activity and fake engagement. In addition, we now breakdown removals by sub-policy. As an example, in Q1, we started publishing the volume of content removed for hateful ideologies and for attacks and slurs, the two sub-policies under our hateful behavior policy, as well as the percentage of content removed proactively, at zero views, and in under 24 hours by sub-policy.

This data is available for download in machine readable formats to support further analysis by researchers, academics, and civil society.

Automated Detection Removals

TikTok uses a combination of people and technology to enforce our Community Guidelines and keep our community safe and welcoming. To do this effectively at scale, we continue to invest in technology-based flagging and moderation. We rely on automated moderation when our systems have a high degree of confidence that content is violative so that we can expeditiously remove violations of our policies. As a result, our overall protective detection efforts have improved.

Keeping our Community Secure

We continue to evolve and adapt our safeguards by investing in automated defenses to detect, block, and remove inauthentic accounts and engagement, and by improving our response speed and efficiency to evolving threats. In the second quarter of 2022, attacks on our systems resulted in an increase in the total volume of fake followers removed. We've implemented measures to hide enforcement actions from malicious actors, preventing them from gaining understanding of our detection capabilities. This led to decreases in spam accounts blocked at sign-up and increases in fake accounts removed.

Fostering a positive advertising experience

TikTok has strict policies to protect users from fake, fraudulent, or misleading content, including ads. Advertiser accounts and ad content are held to these policies and must follow our Community Guidelines, Advertising Guidelines, and Terms of Service. Due to a change in our approach to ad violation enforcement and a strengthening of our account-level enforcement capabilities, total ad removals increased in the first half of 2022. However, in the second quarter of 2022, the total volume of ads removed for violating our advertising policies and guidelines decreased. This is due in part to our efforts to strengthen account-level detection and enforcement strategies, which have helped improve the ads ecosystem and create better experiences for both users and advertisers. Our work to preserve the integrity of our ads ecosystem is never finished, and we will continue to review and further strengthen our systems to combat ads that violate our policies.

Enhancing Our Work to Combat Misinformation

Leveraging machine learning has been especially impactful when it comes to our efforts in countering harmful misinformation. We expanded our capacity to iterate rapidly on our systems given the fast changing nature of misinformation, especially during a crisis or event (e.g. the war in Ukraine or an election). We also improved our ability to detect known misleading audio and imagery to reduce manipulated content on the platform. These investments have yielded quarter over quarter improvements in the enforcement of our integrity and authenticity policies with proactive removal of videos improved from 83.6% in Q1 to 89.1% in Q2; removal of videos at zero views improved from 60.8% in Q1 to 74.7% in Q2; and video removals in under 24 hours improved from 71.9% to 83.9%. To continually improve detecting and removing misinformation, we've made some key investments this year, including:

- continued investment in machine learning models and increased capacity to iterate on these models rapidly given the fast changing nature of misinformation.
- improved detection of known misleading audio and imagery to reduce manipulated content.
- a database of previously fact-checked claims to help misinformation moderators make swift and accurate decisions.
- a proactive detection program with our fact-checkers who flag new and evolving claims they're seeing across the internet. This allows us to look for these claims on our platform and remove violations. Since starting this program in Q1 2022, we identified 33 new misinformation claims, resulting in the removal of 58,000 videos from the platform.

TikTok

Responding to the war in Ukraine

As a platform, we've responded to the war in Ukraine with increased safety and security measures to help ensure people can express themselves and share their experiences while we work aggressively to counter harmful misinformation and other violations of our policies. The war has challenged us to confront a complex and rapidly changing environment. We've increased investments with our fact-checking partners who help assess the accuracy of content. We've continued to evolve our detection methods to help take swifter action against accounts that attempt to scam or mislead our community through unoriginal livestreams and video content. And we began a pilot of our state-controlled media policy by labeling content from some accounts in order to bring important context to viewers. In response to Russia's 'fake news' law, from March 6 we suspended livestreaming and new content to our video service in Russia given the safety implications of this law. Content posted by accounts based outside of Russia is not currently available for distribution in Russia.

From February 24 through the end of the first quarter, March 31, 2022, we took the following steps to safeguard our community:

- Our safety team focused on the Ukraine war removed 41,191 videos, 87% of which violated our policies against harmful misinformation. The vast majority (78%) were identified proactively.
- Our fact-checking partners helped assess 13,738 videos globally.
- We added prompts on 5,600 videos informing viewers that content could not be verified by fact checkers.
- We labeled content from 49 Russian state-controlled media accounts.
- We identified and removed 6 networks and 204 accounts globally for coordinated efforts to influence public opinion and mislead users about their identities.

Our Commitment to Election Integrity

Providing access to authoritative information is an important part of our overall strategy to counter election misinformation. That's why we rolled out out an Elections Center to connect people who engage with election content to authoritative information and sources in more than 45 languages, including English and Spanish. We are committed to promoting digital literacy skills and education, and our in-app center features videos that encourage our community to think critically about content they see online, as well as information about voting in the election.

To ensure that our Elections Center is visible and accessible, we are adding labels to content identified as being related to the 2022 midterm elections as well as content from accounts belonging to governments, politicians, and political parties in the US. These labels allow viewers to click through to our center and get information about the elections in their state. We also provide access on popular elections hashtags, like #elections2022, so that anyone searching for that content are able to easily access the center. We have also begun trialing mandatory verification for accounts belonging to governments, politicians, and political parties through the midterm elections.

TikTok has long prohibited political advertising, including both paid ads on the platform and creators being paid directly to make branded content. We currently do that by prohibiting political content in an ad, and we're also now applying restrictions at an account level. This means accounts belonging to politicians and political parties will automatically have their access to advertising features turned off, which will help us more consistently enforce our existing policy.

Additionally, we will be prohibiting these accounts from accessing other monetization features. Specifically, they will not have access to features like gifting, tipping, and e-commerce, and will be ineligible for our Creator Fund. These changes, along with our existing ban on political advertising, mean that accounts belonging to governments, politicians, and political parties will largely not be able to give or receive money through TikTok's monetization features, or spend money promoting their content. Finally, we have changed our policies to also disallow solicitation for campaign fundraising. That includes content such as a video from a politician asking for donations, or a political party directing people to a donation page on their website.

Providing more ways for our community to enjoy the content they love

Filter Hashtags

We design our recommendation system with safety in mind, since content in someone's For You feed may come from a creator they would prefer not to follow or may relate to an interest they do not share. For instance, certain categories of content may be ineligible for recommendation, and viewers can use our "not interested" feature to automatically skip videos from a creator or that use the same audio. To further empower viewers with ways to customize their viewing experience, we recently rolled out a tool people can use to automatically filter out videos with words or hashtags they don't want to see from their For You or Following feeds - whether because you've just finished a home project and no longer want DIY tutorials or if you want to see fewer dairy or meat recipes as you move to more plant-based meals.

Diversifying Recommendations

Last year we began testing ways to avoid recommending a series of similar content on topics that may be fine as a single video but potentially problematic if viewed repeatedly, such as topics related to dieting, extreme fitness, sadness, and other well-being topics. We've also been testing ways to recognize if our system may inadvertently be recommending a narrower range of content to a viewer. As a result of our tests and iteration in the US, we've improved the viewing experience so viewers now see fewer videos about these topics at a time. We're still iterating on this work given the nuances involved. We're also training our systems to support new languages as we look to expand these tests to more markets.

Content Levels

We have begun working to build a new system to organize content based on thematic maturity. Many people will be familiar with similar systems from their use in the film industry, television, or gaming and we are creating with these in mind while also knowing we need to develop an approach unique to TikTok. In Q3 2022, we began to introduce an early version to help prevent content with overtly mature themes from reaching audiences between ages 13-17. When we detect that a video contains mature or complex themes, a maturity score will be allocated to the video to help prevent those under 18 from viewing it across the TikTok experience. We plan to add new functionality to provide detailed content filtering options for our entire community.

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.

How our policies map to the GARM Brand Safety Floor

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

Note: The above policies reflect our previous Community Guidelines, which we reported against in Q1-Q2 2022. Our guidelines have since been expanded and we will update our mapping to include our new policy areas moving forward.

Question 1: How safe is the platform for consumers?

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

Quarter	Total Removals and Removal Rates		
Q1 2022	**Total Videos Removed: 102,305,516** (Figure represents about 1% of all videos uploaded to TikTok)		
	Videos Removed by Automation: 34,726,592		
	Percentage of videos removed proactively before being reported by a user: 95.1%	**Percentage of videos removed before receiving any views: 90.0%**	**Percentage of videos removed within 24 hours of being posted: 93.7%**
Q2 2022	**Total Videos Removed: 113,809,300** (Figure represents about 1% of all videos uploaded to TikTok)		
	Videos Removed by Automation: 48,011,571		
	Percentage of videos removed proactively before being reported by a user: 95.9%	**Percentage of videos removed before receiving any views: 90.5%**	**Percentage of videos removed within 24 hours of being posted: 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 **Q**2 **202**2	**Q**1 **202**2	**Previous Period Q**4 **2021**	Applicable GARM Categories
Adult nudity and sexual activities	10.7%	11.3%	10.9%	**Adult & Explicit Sexual Content**
Harassment and bullying	5.7%	6.0%	5.7%	**Crime & Harmful acts to individuals and Society, Human Right Violations**
Hateful behavior	1.7%	1.6%	1.5%	**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	21.2%	21.8%	19.5%	**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	0.7%	0.6%	0.6%	**Spam or Harmful Content; Online piracy; Misinformation**
Minor s**afety**	43.7%	41.7%	45.1%	**Adult & Explicit Sexual Content; Crime & Harmful acts to individuals and Society, Human Right Violations**
Suicide, self-harm, and dangerous acts	6.1%	6.7%	7.4%	**Debated Sensitive Social Issue; Crime & Harmful acts to individuals and Society, Human Right Violations; Death, Injury or Military Conflict**
Violent and graphic content	9.3%	9.6%	8.5%	**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	0.9%	0.7%	0.8%	**Terrorism**

Question 2: How safe is the platform for advertisers?

Not submitted

Relevant Policy	Latest Period	Previous Period	Commentary
Adult nudity and sexual activities			
Harassment and bullying			
Hateful behavior			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.
Illegal activities and regulated goods			
Integrity and authenticity			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.
Minor safety			
Suicide, self-harm, and dangerous acts			
Violent and graphic content			
Violent extremism			

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

Next Best Measure: Removal rates by policy

T	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	Applicable GARM Categories
Adult nudity and sexual activities	90.2%	80.3%	90.1%	89.4%	78.3%	88.9%	90.3%	79.6%	90.5%	**Adult & Explicit Sexual Content**
Harassment and bullying	82.4%	71.4%	83.5%	78.9%	69.4%	83.9%	77.2%	66.9%	84.5%	**Crime & Harmful acts to individuals and Society, Human Right Violations**
Hateful behavior	81.1%	72.4%	83.5%	77.0%	68.3%	82.2%	76%	65.2%	82.5%	**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%	93.1%	94.9%	97.1%	93.8%	95.6%	96.8%	93%	95.7%	**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	89.1%	74.7%	83.9%	83.6%	60.8%	71.9%	85.5%	67%	75.7%	**Spam or Harmful Content; Online piracy; Misinformation**
Minor safety	98.4%	95.4%	95.7%	98.1%	95.5%	95.9%	98.2%	95.6%	96.20%	**Adult & Explicit Sexual Content; Crime & Harmful acts to individuals and Society, Human Right Violations**
Suicide, self-harm, and dangerous acts	96.2%	85.4%	88.5%	95.3%	86.1%	90.4%	96%	87.6%	92.7%	**Debated Sensitive Social Issue; Crime & Harmful acts to individuals and Society, Human Right Violations; Death, Injury or Military Conflict**
Violent and graphic content	97.3%	91.5%	94.0%	96.4%	89.9%	94.3%	96.1%	89.9%	94.9%	**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	92.8%	82.2%	85.6%	91.4%	83.9%	88.4%	92.2%	83.5%	89.5%	**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.

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

Next Best Measure: Removal rates by sub-policy.

TikTok Issue Policy	-	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	87.4%	74.4%	85.8%	87.7%	74.1%	85.4%	N/A		
	Sexual exploitation	81.0%	64.7%	85.7%	80.4%	64.7%	87.2%			
Harassment and bullying	Abusive behavior	81.2%	70.3%	82.7%	78.2%	68.9%	83.6%	N/A		
	Sexual harassment	73.3%	53.7%	74.7%	68.9%	52.2%	75.8%			
	Threats of hacking, doxxing, and blackmail	87.8%	79.4%	81.3%	87.5%	81.4%	86.6%			
Hateful behavior	Attacks and slurs on the basis of protected attributes	85.0%	75.9%	84.5%	81.7%	72.0%	83.5%	N/A		
	Hateful ideology	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.

TikTok Issue Policy		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**	87.4%	66.1%	80.2%	82.8%	62.4%	80.3%	N/A		
	Drugs, controlled substances, alcohol, and tobacco	94.4%	83.5%	86.9%	94.2%	86.6%	89.5%			
	Frauds and scams	83.1%	62.8%	81.0%	79.2%	68.9%	83.9%			
	Gambling	95.2%	77.5%	85.9%	94.3%	82.5%	88.8%			
	Privacy, personal data, and personally identifiable information	98.8%	94.3%	93.5%	98.4%	95.4%	95.2%			
	Weapons	97.4%	92.6%	90.2%	97.2%	93.9%	93.0%			
Integrity and **authenticity**	**Harmful Misinformation**	70.0%	39.1%	57.5%	66.1%	37.6%	50.6%	N/A		
	Spam and fake engagement	82.9%	59.9%	76.6%	84.4%	66.1%	81.3%			

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.

TikTok Issue Policy	-	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	96.7%	88.4%	90.0%	96.6%	90.6%	92.7%	N/A		
	Harmful activities by minors	97.1%	90.7%	88.0%	96.8%	92.9%	91.3%			
	Nudity and sexual activity involving minors	96.9%	91.4%	89.9%	96.8%	92.6%	91.8%			
	Physical and psychological harm of a minor	96.7%	87.8%	86.5%	96.2%	88.2%	90.1%			
	Sexual exploitation of minors	93.2%	85.8%	90.7%	90.6%	82.5%	90.3%			
Suicide, self-harm, and dangerous acts	Dangerous acts and challenges	95.0%	78.8%	84.5%	94.4%	82.6%	88.8%	N/A		
	Disordered eating	86.7%	75.2%	82.5%	82.0%	70.9%	82.2%			
	Suicide and self-harm	97.1%	93.4%	91.5%	97.1%	94.0%	94.0%			

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.

TikTok Issue Policy	-	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 g**raphic** c**ontent**	**Violent and graphic content**	95.9%	86.9%	89.7%	95.0%	86.4%	91.8%	N/A		
Violent extremism	**Threats and incitement to violence**	84.7%	75.6%	84.3%	80.9%	69.5%	82.4%	N/A		
	Violent extremist organizations and individuals	93.0%	85.9%	88.8%	93.3%	87.0%	89.8%			

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?

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	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	Commentary
Adult nudity and sexual activities	We removed a total of **59,430,082** accounts for violating Community Guidelines or Terms of Service.	20,575,056	33,632,058	5,222,968	We removed a total of **44,438,988** accounts for violating Community Guidelines or Terms of Service.	20,219,476	20,890,519	3,328,993	We removed a total of **24,107,316** accounts for violating Community Guidelines or Terms of Service.	15,383,165	6,077,046	2,647,105	Account Removals represents figure across all Community Guidelines.
Harassment and bullying													
Hateful behavior													
Illegal activities and regulated goods													
Integrity and authenticity													
Minor safety													
Suicide, self-harm, and dangerous acts													
Violent and graphic content													
Violent extremism													

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

Authorized Metric: Removal of Violating Accounts

Adult nudity and sexual activities				
Harassment and bullying				
Hateful behavior	We reinstated **5,896,218** videos after they were appealed	We reinstated **5,025,536** videos after they were appealed	We reinstated **4,727,382** videos after they were appealed	Content reinstatement represents figure across all Community Guidelines
Illegal activities and regulated goods				
Integrity and authenticity				
Minor safety				
Suicide, self-harm, and dangerous acts				
Violent and graphic content				
Violent extremism				

Pinterest

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. And we're thoughtful about where ads do show up. For instance, we don't monetize search terms related to the coronavirus pandemic.

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.

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.

***

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 biannual transparency report in 2013, and in 2021 we expanded the report to include new information. Now, our bi-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 climate misinformation, a policy we rolled out in April 2022 to keep false and misleading claims around climate change off the platform.

Our latest transparency report includes data form Q1 2022 (January–March 2022) and Q2 2022 (April–June 2022). During this reporting period, we continued to support the health and wellbeing of our community by expanding compassionate search to 11 more countries and launching hair pattern search in Europe and Latin America. We stood with Ukraine by prioritizing humanitarian aid for people affected and are continuing to work to keep our platform safe from disinformation.

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.

Pinterest

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.

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 for this type of content.

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 Q1 and Q2 2022.

Question 1: How safe is the platform for consumers?
Next best measure: Reach[1] of Pins deactivated for violating policy

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

Pinterest Policy[2]	Latest Period								Previous Period							
	Q2 2022				Q1 2022				Q4 2021				Q3 2021			
	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people
Adult content	77%	19%	3%	1%	78%	18%	3%	1%	88%	10%	2%	0.6%	82%	14%	2%	1%
Adult sexual services	82%	14%	0.5%	0.9%	95%	5%	0.5%	0.1%	89%	9%	2%	0.6%	0.4%[3]	47%	29%	24%
Child sexual exploitation (CSE)[4]	61%	30%	6%	3%	63%	28%	6%	3%	83%	13%	3%	2%	72%	21%	4%	2%
Civic misinformation	46%	47%	5%	3%	20%	67%	5%	8%	99.5%	0.4%	0.06%	0.04%	10%	78%	7%	5%
Climate Misinformation	13%	81%	4%	2%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Conspiracy theories	84%	14%	1%	1%	57%	33%	5%	5%	89%	8%	1%	2%	93%	6%	0.3%	0.1%
Copyright[5]	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	64%	31%	3%	2%	72%	21%	4%	3%	76%	16%	5%	3%	95%	4%	0.9%	0.3%
Graphic violence and threats	67%	28%	2%	4%	65%	29%	4%	3%	59%	13%	10%	19%	52%	29%	8%	11%
Harassment and criticism	58%	35%	4%	2%	74%	21%	3%	3%	69%	26%	3%	2%	86%	12%	2%	1%
Hateful activities	62%	26%	4%	8%	63%	22%	6%	9%	97%	0.8%	0.7%	1%	56%	19%	11%	14%
Medical misinformation	91%	8%	0.3%	0.4%	85%	13%	0.7%	1%	76%	9%	5%	9%	75%	20%	4%	1%
Self-injury and harmful behavior	93%	6%	0.6%	0.4%	71%	23%	3%	3%	97%	2%	0.3%	0.3%	83%	12%	3%	2%
Spam	89%	10%	0.7%	0.3%	60%	34%	4%	1%	75%	19%	4%	1%	63%	24%	9%	4%
Trademark[5]	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

[1] 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.

[2] Reach for Civic misinformation in Q4 2021 does not include Pins deactivated in the course of one-time sweeps that we later determined to be false positives and subsequently reinstated. See commentary for more details.

[3] Of the 225 Pins deactivated in Q3 2021 for violating our Adult sexual services policy, 171 were seen by fewer than 100 users in that reporting period.

[4] 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.

[5] We do not currently report on reach for Copyright or Trademark.

Question 2: How safe is the platform for advertisers?
Next best measure: Reach[1] of Pins deactivated for violating policy

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

| | Latest Period | | | | | | | | Previous Period | | | | | | | |
| | Q2 2022 | | | | Q1 2022 | | | | Q4 2021 | | | | Q3 2021 | | | |
Pinterest Policy[2]	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people
Adult content	77%	19%	3%	1%	78%	18%	3%	1%	88%	10%	2%	0.6%	82%	14%	2%	1%
Adult sexual services	82%	14%	0.5%	0.9%	95%	5%	0.5%	0.1%	89%	9%	2%	0.6%	0.4%[3]	47%	29%	24%
Child sexual exploitation (CSE)[4]	61%	30%	6%	3%	63%	28%	6%	3%	83%	13%	3%	2%	72%	21%	4%	2%
Civic misinformation	46%	47%	5%	3%	20%	67%	5%	8%	99.5%	0.4%	0.06%	0.04%	10%	78%	7%	5%
Climate Misinformation	13%	81%	4%	2%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Conspiracy theories	84%	14%	1%	1%	57%	33%	5%	5%	89%	8%	1%	2%	93%	6%	0.3%	0.1%
Copyright[5]	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	64%	31%	3%	2%	72%	21%	4%	3%	76%	16%	5%	3%	95%	4%	0.9%	0.3%
Graphic violence and threats	67%	28%	2%	4%	65%	29%	4%	3%	59%	13%	10%	19%	52%	29%	8%	11%
Harassment and criticism	58%	35%	4%	2%	74%	21%	3%	3%	69%	26%	3%	2%	86%	12%	2%	1%
Hateful activities	62%	26%	4%	8%	63%	22%	6%	9%	97%	0.8%	0.7%	1%	56%	19%	11%	14%
Medical misinformation	91%	8%	0.3%	0.4%	85%	13%	0.7%	1%	76%	9%	5%	9%	75%	20%	4%	1%
Self-injury and harmful behavior	93%	6%	0.6%	0.4%	71%	23%	3%	3%	97%	2%	0.3%	0.3%	83%	12%	3%	2%
Spam	89%	10%	0.7%	0.3%	60%	34%	4%	1%	75%	19%	4%	1%	63%	24%	9%	4%
Trademark[5]	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

[1] 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.

[2] Reach for Civic misinformation in Q4 2021 does not include Pins deactivated in the course of one-time sweeps that we later determined to be false positives and subsequently reinstated. See commentary for more details.

[3] Of the 225 Pins deactivated in Q3 2021 for violating our Adult sexual services policy, 171 were seen by fewer than 100 users in that reporting period.

[4] 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.

[5] We do not currently report on reach for Copyright or Trademark.

Question 3: How safe is the platform for advertisers?
Next best measure: Reach[1] of Pins deactivated for violating policy

	Latest Period								Previous Period							
Pinterest Policy[2]	**Q2 2022**				**Q1 2022**				**Q4 2021**				**Q3 2021**			
	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people	Seen by 0 people	Seen by <10 people	Seen by 10-100 people	Seen by 100+ people
Adult content	77%	19%	3%	1%	78%	18%	3%	1%	88%	10%	2%	0.6%	82%	14%	2%	1%
Adult sexual services	82%	14%	0.5%	0.9%	95%	5%	0.5%	0.1%	89%	9%	2%	0.6%	0.4%[3]	47%	29%	24%
Child sexual exploitation (CSE)[4]	61%	30%	6%	3%	63%	28%	6%	3%	83%	13%	3%	2%	72%	21%	4%	2%
Civic misinformation	46%	47%	5%	3%	20%	67%	5%	8%	99.5%	0.4%	0.06%	0.04%	10%	78%	7%	5%
Climate Misinformation	13%	81%	4%	2%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Conspiracy theories	84%	14%	1%	1%	57%	33%	5%	5%	89%	8%	1%	2%	93%	6%	0.3%	0.1%
Copyright[5]	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	64%	31%	3%	2%	72%	21%	4%	3%	76%	16%	5%	3%	95%	4%	0.9%	0.3%
Graphic violence and threats	67%	28%	2%	4%	65%	29%	4%	3%	59%	13%	10%	19%	52%	29%	8%	11%
Harassment and criticism	58%	35%	4%	2%	74%	21%	3%	3%	69%	26%	3%	2%	86%	12%	2%	1%
Hateful activities	62%	26%	4%	8%	63%	22%	6%	9%	97%	0.8%	0.7%	1%	56%	19%	11%	14%
Medical misinformation	91%	8%	0.3%	0.4%	85%	13%	0.7%	1%	76%	9%	5%	9%	75%	20%	4%	1%
Self-injury and harmful behavior	93%	6%	0.6%	0.4%	71%	23%	3%	3%	97%	2%	0.3%	0.3%	83%	12%	3%	2%
Spam	89%	10%	0.7%	0.3%	60%	34%	4%	1%	75%	19%	4%	1%	63%	24%	9%	4%
Trademark[5]	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

[1] 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.

[2] Reach for Civic misinformation in Q4 2021 does not include Pins deactivated in the course of one-time sweeps that we later determined to be false positives and subsequently reinstated. See commentary for more details.

[3] Of the 225 Pins deactivated in Q3 2021 for violating our Adult sexual services policy, 171 were seen by fewer than 100 users in that reporting period.

[4] 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.

[5] We do not currently report on reach for Copyright or Trademark.

Question 3: How effective is the platform in enforcing safety policy?
Authorized Metric: Distinct images deactivated[1], Pins deactivated[2], Boards deactivated[3], Accounts deactivated[4]

Violating content deactivated by Pinterest.

| | Latest Period | | | | | | | | Previous Period | | | | | | | |
| | Q2 2022 | | | | Q1 2022 | | | | Q4 2021 | | | | Q3 2021 | | | |
Pinterest Policy	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	1,058,729	27,606,227	43,757	6,571	975,774	22,309,237	47,965	7,699	1,108,006	30,695,545	44,155	6,769	1,279,861	42,968,305	61,114	8,653
Adult sexual services	21,746	51,936	146	156	6,052	208,296	129	117	1,610	26,703	120	121	222	225	151	152
Child sexual exploitation (CSE)[5]	9,085	712,295	1,162	37,694	2,499	300,003	492	10,743	2,545	104,029	578	17,423	2,362	262,164	862	28,289
Civic misinformation	1,541	2,782	33	9	712	840	42	10	779	48,741	23	0	743	889	92	8
Climate Misinformation	1,120	1,218	132	7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Conspiracy theories	3,857	14,312	311	19	2,579	5,039	223	19	2,411	4,507	237	11	8,374	536,455	429	39
Copyright	41,667	12,898,085	0	381	63,817	13,995,413	0	380	35,966	10,835,478	0	292	51,833	12,520,214	0	714
Dangerous goods and activities	12,253	81,286	583	245	11,308	55,985	603	222	14,736	31,186	591	177	44,277	1,017,444	712	223
Graphic violence and threats	12,845	96,854	633	60	9,662	124,008	800	120	3,435	6,085	769	89	4,026	18,501	1,085	106
Harassment and criticism	9,672	232,858	1,356	266	3,128	61,237	931	197	4,318	210,880	803	151	4,257	244,688	851	271
Hateful activities	5,774	40,455	705	76	4,016	27,585	726	105	3,848	107,295	409	36	3,482	9,805	1,758	107
Medical misinformation	3,906	262,909	299	4	4,426	113,195	513	17	3,564	6,370	491	14	4,237	138,491	383	10
Self-injury and harmful behavior	8,494	1,232,828	14,667	29	6,574	107,059	16,144	61	16,761	395,682	82,518	1,047	4,156	85,328	575	37
Spam	37,398	259,996	10,543	1,680,646	35,243	74,941	0	1,477,695	61,834	121,639	0	1,736,742	81,753	233,303	0	2,435,683
Trademark	27,125	34,366	164	241	27,859	36,215	241	205	25,366	33,181	150	135	29,557	36,400	82	128

[1] 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.
[2] Does not include Pins that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.
[3] 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.
[4] When policy-violating accounts are removed, all Pins and boards belonging to those accounts are removed as well.
[5] 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[1], Pins deactivated[2], Boards deactivated[3], Accounts deactivated[4]

Violating content deactivated by Pinterest

Pinterest policy	Commentary
Adult content	We deactivated fewer Pins for violating this policy in Q1 2022 than in Q2 2022. Of the Pins we deactivated in Q1 and Q2, 99% were seen by less than 100 users in that reporting period.
Adult sexual services	We deactivated more Pins for violating our adult sexual services policy in Q1 2022 than in Q2 2022. This was the result of a hybrid deactivation of a small handful of images, which account for almost two-thirds of Pins deactivated in Q1 for violating this policy. Of the Pins we deactivated in Q1 and Q2, at least 99% were seen by less than 100 users in that reporting period.
Child sexual exploitation (CSE)	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. 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.
Civic misinformation	We determined that hybrid deactivations based on one Pin, which had been deactivated for reasons other than the image it showed, resulted in the incorrect deactivation of its almost 24,000 machine-identified matching Pins, and we reinstated the content after spotting the error. 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.
Climate Misinformation	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 may harm the public's well-being, safety or trust, including things like 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.
Conspiracy theories	We deactivated fewer Pins for violating this policy in Q1 2022 than in Q2 2022. Of the Pins we deactivated in Q2, 99% were seen by less than 100 users in that reporting period.
Copyright	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. In cases where rightsholders 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.
Dangerous goods and activities	We deactivated fewer Pins for violating this policy in Q1 2022 than in Q2 2022. Of the Pins we deactivated in Q2, 98% were seen by less than 100 users in that reporting period.
Graphic violence and threats	In Q1 2022, we saw a large increase in Pins deactivated compared to Q4 2021, in part because of the increase in content related to the war in Ukraine. Of the Pins we deactivated in Q1 2022, 97% were seen by fewer than 100 users in that reporting period.

[1] 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.
[2] Does not include Pins that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.
[3] 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.
[4] When policy-violating accounts are removed, all Pins and boards belonging to those accounts are removed as well.

Authorized Metric: Distinct images deactivated[1], Pins deactivated[2], Boards deactivated[3], Accounts deactivated[4]

Violating content deactivated by Pinterest

Pinterest policy	Commentary
Harassment and criticism	We deactivated fewer Pins for violating this policy in Q1 2022 than in Q2 2022. Of the Pins we deactivated in Q2, 98% were seen by fewer than 100 users in this reporting period.
Hateful activities	In Q4 2021, we performed a sweeping cleanup across the platform for content violating our conspiracy theories policy that generated a temporary spike in Pins deactivated. This rise is not due to any known increase of violative content. We deactivated fewer Pins for violating this policy in Q1 2022 than in Q2 2022. Of the Pins we deactivated in Q2, 92% were seen by fewer than 100 users in this reporting period.
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 fewer Pins for violating this policy in Q1 2022 than in Q2 2022. Of the Pins we deactivated in Q2, more than 99% were seen by fewer than 100 users in this reporting period.
Self-injury and harmful behavior	We continued investing in work to improve content moderation for self-harm content and providing compassionate support for those in need. As a result, we saw a large increase in Pins deactivated in Q2 2022. Of the Pins we deactivated in Q2, 99% were seen by fewer than 100 users in this 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. 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 Q1 2022 than in Q2 2022. Of the Pins we deactivated in Q2, more than 99% were seen by fewer than 100 users in this 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.

[1] 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.
[2] Does not include Pins that were deactivated because they were on a board that was deactivated or belonged to a user that was deactivated.
[3] 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.
[4] 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

	Latest Period						Previous Period					
Pinterest Policy	**Q2 2022**			**Q1 2022**			**Q4 2021**			**Q3 2021**		
	Automated[1]	Manual[2]	Hybrid[3]	Automated	Manual	Hybrid	Automated[1]	Manual[2]	Hybrid[3]	Automated	Manual	Hybrid
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%	100%	0%
Child sexual exploitation	0%	1%	99%	0%	<1%	>99%	0%	2%	98%	0%	<1%	>99%
Civic misinformation	0%	69%	31%	0%	90%	10%	0%	<1%	>99%	0%	87%	13%
Climate Misinformation	0%	>99%	<1%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Conspiracy theories	0%	77%	23%	0%	67%	33%	0%	90%	10%	0%	<1%	>99%
Copyright	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%
Dangerous goods and activities	7%	12%	81%	12%	16%	72%	14%	41%	45%	<1%	2%	98%
Graphic violence and threats	0%	27%	73%	<1%	9%	91%	0%	62%	38%	0%	48%	53%
Harassment and criticism	0%	4%	96%	0%	5%	95%	0%	2%	98%	0%	1%	99%
Hateful activities	0%	15%	85%	0%	16%	84%	0%	5%	95%	0%	46%	54%
Medical misinformation	2%	2%	97%	4%	3%	93%	72%	28%	<1%	2%	25%	74%
Self-injury and harmful behavior	0%	<1%	>99%	0%	6%	94%	0%	1%	99%	0%	5%	95%
Spam	>99%	<1%	0%	>99%	<1%	<1%	>99%	<1%	0%	>99%	<1%	0%
Trademark	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%

[1] 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.

[2] 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.

[3] 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 prevalence 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

| | Latest Period | | | | Previous Period | | | |
| | Q2 2022 | | Q1 2022 | | Q4 2021 | | Q3 2021 | |
Pinterest Policy	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated
Adult content	1,674	1,070	2,104	1,229	1,495	774	1,923	1,119
Adult sexual services	4	1	9	3	16	1	13	0
Child sexual exploitation	7,467	5,971	2,164	1,169	3,110	2,120	5,718	4,305
Civic misinformation	4	2	4	1	1	1	3	3
Climate Misinformation	0	0	N/A	N/A	N/A	N/A	N/A	N/A
Conspiracy theories	37	8	6	4	13	2	19	4
Copyright[1]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Dangerous goods and activities	11	3	12	2	4	1	14	3
Graphic violence and threats	25	14	25	9	19	3	33	18
Harassment and criticism	55	38	28	16	28	15	22	13
Hateful activities	33	12	31	6	10	5	35	21
Medical misinformation	1	1	4	1	4	3	2	0
Self-injury and harmful behavior	21	10	30	9	166	133	10	9
Spam	73,639	55,930	98,156	79,469	103,257	79,054	101,832	77,936
Trademark[1]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

[1] We do not currently report on account appeals and reinstatements for Copyright or Trademark.

Snapchat

At Snap, our core underlying belief is in the need to build a safe platform for our community, and for the world. That is the goal that drives many of our unique design and policy choices. We built Snapchat around the camera because we wanted to create a new way to give people a way to express their full experiences, with their real friends.

During this reporting period, seven percent of all content we enforced against globally, and 10 percent of all content we enforced against in the U.S., involved drug-related violations. Globally, the median turnaround time we took action to enforce against these accounts was within 10 minutes of receiving a report.

Over the past year, we have been deeply focused on combating the rise of illicit drug activity as part of the larger growing fentanyl and opioid epidemic across the U.S. We take a holistic approach that includes deploying tools that proactively detect drug-related content, working with law enforcement to support their investigations, and providing in-app information and support to Snapchatters through our fentanyl-related education portal, Heads Up. Heads Up surfaces resources from expert organizations when Snapchatters search for a range of drug-related terms and their derivatives, which we also block. As a result of these ongoing efforts, the vast majority of drug-related content we uncover is proactively detected by our machine learning and artificial intelligence technology, and we will continue working to make as much progress as possible to eradicate drug dealers from our platform.

We have also created a new suicide and self-harm category to share the total number of content and account reports that we received and took action on when our Trust & Safety teams determined that a Snapchatter may be in crisis. We care deeply about the mental health and wellbeing of Snapchatters and believe we have a duty to support our community in these difficult moments.

In addition to these new elements in our latest Transparency Report, our data shows that we saw a reduction in two key areas: Violative View Rate (VVR) and the number of accounts we enforced that attempted to spread hate speech, violence, or harm. Our current Violative View Rate is (VVR) 0.08 percent. This means that out of every 10,000 Snap and Story views on Snapchat, eight contained content that violated our Community Guidelines. This is an improvement from our last reporting cycle, during which our VVR was 0.10 percent.

While the fundamental architecture of Snapchat protects against the ability for harmful content to go viral, we continue to be vigilant and improve our human moderation. As a result, we have improved the median enforcement turnaround time by 25 percent for hate speech and eight percent for threats and violence or harm to 12 minutes in both categories.

We believe it's our responsibility to keep our community safe on Snapchat and we are constantly evaluating how we can continue to strengthen our comprehensive efforts to do that. Our work here is never done, but we will continue communicating updates about our progress and we are grateful to our many partners that regularly help us improve.

Question 1: How safe is the platform for consumers?

Authorized Metric: Violation view rate

An estimate of the percentage of story views that violated our community guidelines in a given reporting period

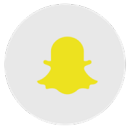| GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Sexually Explicit Content			We prohibit accounts that promote or distribute pornographic content. We report child sexual exploitation to authorities. Never post, save, or send nude or sexually explicit content involving anyone under the age of 18 — even of yourself. Never ask a minor to send explicit imagery or chats. Breastfeeding and other depictions of nudity in certain non-sexual contexts may be permitted.
Arms & Ammunition	Regulated Goods	During the reporting period, we saw a Violative View Rate (VVR) of 0.08 percent	During the reporting period, we saw a Violative View Rate (VVR) of 0.08 percent	Don't use Snapchat for any illegal activities — including to buy or sell illegal drugs, contraband, counterfeit goods, or illegal weapons. We prohibit the promotion and use of certain regulated goods, as well as the depiction or promotion of criminal activities.
Crime & Harmful acts to individuals and Society, Human Right Violations	Threatening / Violence / Harm			Encouraging violence or dangerous behavior is prohibited — never threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury or eating disorders.
Death, Injury or Military Conflict	Threatening / Violence / Harm			Encouraging violence or dangerous behavior is prohibited — never threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury or eating disorders.
Online piracy	Spam			Pretending to be someone you're not — this includes your friends, celebrities, brands, or other organizations — or attempting to deceive people about who you are is not allowed. We prohibit spam and other deceptive practices, including manipulating content for misleading purposes or to imitate Snapchat content formats.

Question 1: How safe is the platform for consumers?

Authorized Metric: Violation view rate

An estimate of the percentage of story views that violated our community guidelines in a given reporting period

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Hate speech & acts of aggression	Hate Speech	During the reporting period, we saw a Violative View Rate (VVR) of 0.08 percent	During the reporting period, we saw a Violative View Rate (VVR) of 0.08 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
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	N/A			As standalone, this does 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.
Illegal Drugs / Tobacco / e-cigarettes / Vaping / Alcohol	Regulated Goods			Don't use Snapchat for any illegal activities — including to buy or sell illegal drugs, contraband, counterfeit goods, or illegal weapons. We prohibit the promotion and use of certain regulated goods, as well as the depiction or promotion of criminal activities.
Spam or Harmful Content	Spam			Pretending to be someone you're not — this includes your friends, celebrities, brands, or other organizations — or attempting to deceive people about who you are is not allowed. We prohibit spam and other deceptive practices, including manipulating content for misleading purposes or to imitate Snapchat content formats.
Terrorism	Terrorism			Terrorist organizations and hate groups are prohibited from using our platform and we have no tolerance for content that advocates or advances violent extremism or terrorism.
Debated Sensitive Social Issue	N/A			We do not report on this category, but Snap is actively involved in discussions with GARM and member platforms to break out subjects within this category, notably misinformation / disinformation.

Question 1: How safe is the platform for consumers?

Authorized Metric: Violation view rate

An estimate of the percentage of story views that violated our community guidelines in a given reporting period

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Sexually Explicit Content			We prohibit accounts that promote or distribute pornographic content. We report child sexual exploitation to authorities. Never post, save, or send nude or sexually explicit content involving anyone under the age of 18 — even of yourself. Never ask a minor to send explicit imagery or chats. Breastfeeding and other depictions of nudity in certain non-sexual contexts may be permitted.
Arms & Ammunition	Regulated Goods	During the reporting period, we saw a Violative View Rate (VVR) of 0.08 percent	During the reporting period, we saw a Violative View Rate (VVR) of 0.08 percent	Don't use Snapchat for any illegal activities — including to buy or sell illegal drugs, contraband, counterfeit goods, or illegal weapons. We prohibit the promotion and use of certain regulated goods, as well as the depiction or promotion of criminal activities.
Crime & Harmful acts to individuals and Society, Human Right Violations	Threatening / Violence / Harm			Encouraging violence or dangerous behavior is prohibited — never threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury or eating disorders.
Death, Injury or Military Conflict	Threatening / Violence / Harm			Encouraging violence or dangerous behavior is prohibited — never threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury or eating disorders.
Online piracy	Spam			Pretending to be someone you're not — this includes your friends, celebrities, brands, or other organizations — or attempting to deceive people about who you are is not allowed. We prohibit spam and other deceptive practices, including manipulating content for misleading purposes or to imitate Snapchat content formats.

Question 2: How safe is the platform for advertisers?

Authorized Metric: Violation view rate

An estimate of the percentage of story views that violated our community guidelines in a given reporting period

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Hate speech & acts of aggression	Hate Speech			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
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	N/A	During the reporting period, we saw a Violative View Rate (VVR) of 0.08 percent	During the reporting period, we saw a Violative View Rate (VVR) of 0.08 percent	As standalone, this does 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.
Illegal Drugs / Tobacco / e-cigarettes / Vaping / Alcohol	Regulated Goods			Don't use Snapchat for any illegal activities — including to buy or sell illegal drugs, contraband, counterfeit goods, or illegal weapons. We prohibit the promotion and use of certain regulated goods, as well as the depiction or promotion of criminal activities.
Spam or Harmful Content	Spam			Pretending to be someone you're not — this includes your friends, celebrities, brands, or other organizations — or attempting to deceive people about who you are is not allowed. We prohibit spam and other deceptive practices, including manipulating content for misleading purposes or to imitate Snapchat content formats.
Terrorism	Terrorism			Terrorist organizations and hate groups are prohibited from using our platform and we have no tolerance for content that advocates or advances violent extremism or terrorism.
Debated Sensitive Social Issue	N/A			We do not report on this category, but Snap is actively involved in discussions with GARM and member platforms to break out subjects within this category, notably misinformation / disinformation.

Next Best Measure: Content Actioned, Actors Actioned

Content removed by Snap - Users removed by Snap

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Content Actioned	**Actors Actioned**	**Content Actioned**	**Actors Actioned**	
Adult & Explicit Sexual Content	Sexually Explicit Content	**4,869,272**	**1,716,547**	**4,783,518**	**1,441,208**	We prohibit accounts that promote or distribute pornographic content. We report child sexual exploitation to authorities. Never post, save, or send nude or sexually explicit content involving anyone under the age of 18 — even of yourself. Never ask a minor to send explicit imagery or chats. Breastfeeding and other depictions of nudity in certain non-sexual contexts may be permitted.
Arms & Ammunition	Weapons	**28,706**	**21,310**	**620,083**	**274,883**	As part of our ongoing focus on improving our transparency reports, we are introducing several new elements to this report. For this installment and going forward, we are breaking out drugs, weapons and regulated goods into their own categories, which will provide additional detail about their prevalence and our enforcement efforts.
Crime & Harmful acts to individuals and Society, Human Right Violations	Threatening / Violence / Harm	**232,565**	**159,214**	**465,422**	**288,091**	Encouraging violence or dangerous behavior is prohibited — never threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury or eating disorders.
Death, Injury or Military Conflict	Threatening / Violence / Harm	**232,565**	**159,214**	**465,422**	**288,091**	Encouraging violence or dangerous behavior is prohibited — never threaten to harm a person, a group of people, or someone's property. Snaps of gratuitous or graphic violence are not allowed. We don't allow the glorification of self-harm, including the promotion of self-injury or eating disorders.
Online piracy	Spam	**153,621**	**110,102**	**243,729**	**120,898**	Pretending to be someone you're not — this includes your friends, celebrities, brands, or other organizations — or attempting to deceive people about who you are is not allowed. We prohibit spam and other deceptive practices, including manipulating content for misleading purposes or to imitate Snapchat content formats.

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

Next Best Measure: Content Actioned, Actors Actioned

Content removed by Snap - Users removed by Snap

Some individual Snap categories encompass multiple GARM categories (example: GARM'S Online Piracy and Spam categories both roll up under "Spam" in Snap's TR). 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 Policy	Latest Period		Previous Period		Commentary
		Content Actioned	Actors Actioned	Content Actioned	Actors Actioned	
Hate speech & acts of aggression	Hate Speech	93,341	63,767	121,639	92,314	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
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust						As standalone, this does 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.
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Drugs	428,311	278,304	620,083	274,883	As part of our ongoing focus on improving our transparency reports, we are introducing several new elements to this report. For this installment and going forward, we are breaking out drugs, weapons and regulated goods into their own categories, which will provide additional detail about their prevalence and our enforcement efforts.
Spam or Harmful Content	Spam	153,621	110,102	243,729	110,102	Pretending to be someone you're not — this includes your friends, celebrities, brands, or other organizations — or attempting to deceive people about who you are is not allowed. We prohibit spam and other deceptive practices, including manipulating content for misleading purposes or to imitate Snapchat content formats.
Terrorism	Terrorism	14,613	22	119,134	5	Terrorist organizations and hate groups are prohibited from using our platform and we have no tolerance for content that advocates or advances violent extremism or terrorism.
Debated Sensitive Social Issue	N/A					We do not report on this category, but Snap is actively involved in discussions with GARM and member platforms to break out subjects within this category, notably misinformation / disinformation.

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

Next Best Measure: Content Actioned, Actors Actioned

Content removed by Snap - Users removed by Snap

GARM Category	Relevant Policy	Latest Period	Previous	Commentary
Adult & Explicit Sexual Content	Child Sexual Exploitation and Abuse	198,109	119,134	In the second half of 2021, we proactively detected and actioned 88 percent of the total CSAM violations reported here.
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	Terrorist & Violent Extremist Content	22	5	At Snap, we remove terrorist and violent extremism content reported through multiple channels. These include allowing users to report terrorist and violent extremist content through our in-app reporting menu, and we work closely with law enforcement to address terrorism and violent extremism content that may appear on Snap.
Debated Sensitive Social Issue				

Global Alliance for Responsible Media

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

Next Best Measure: Content Actioned, Actors Actioned

Content removed by Snap - Users removed by Snap

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 currently report on this metric								
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?

Not applicable to Snap

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Content Appealed	Content Reinstated	Content Appealed	Content Reinstated	
Adult & Explicit Sexual Content						
Arms & Ammunition						
Crime & Harmful acts to individuals and Society, Human Right Violations						
Death, Injury or Military Conflict						
Online piracy						
Hate speech & acts of aggression			Snap does not currently offer an appeals process, and therefore does not report on this metric			
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						

Twitch

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 creators, 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, to safety livestreams, 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.

H1 2022 Overview
We are constantly investing in tooling to ensure all of our users are authentically, safely, and meaningfully engaging with each other. In H1 2022, Twitch updated its existing reporting tool to include more report options and a more intuitive interface so users can better file reports of violative content. By empowering our users to report content more quickly and efficiently, we can take action against violative content even faster.

In the first half of 2022, we made revisions to our Community Guidelines, introducing new standards aimed at reducing the risk of severe harm to our community, and clarifying existing policies. We updated our guidelines for account usernames and display names to prohibit names that reference hard drugs or sexual content. Finally, we made updates to our self-harm guidelines to prevent glorification of self-destructive behavior such as eating disorders.

We're continuing to iterate our portfolio of tools offered to users. AutoMod, when enabled, pre-screens chat messages in 17 languages and holds messages that contain content detected as risk, preventing them from being visible in chat unless they are approved by a moderator. Suspicious User Detection uses machine learning to predict the likelihood an account is attempting to evade a channel-level ban. Mod View brings all of our tools together in a customizable channel interface that provides moderators with a set of 'widgets' for moderation tasks.

We also invested in giving our streamers and viewers a better appeals portal so they can review recent enforcement history, select a specific enforcement to appeal, and view the status of previous appeals. It offers a new level of clarity around the enforcements they received helping reduce user confusion. Along with external tools, we improved our internal tooling to better process and review submitted appeals. This helps us make consistent decisions and allows us to process a higher volume of appeals to ensure everyone who submits an appeal can get a timely and accurate response.

Methodology
Our Community Guidelines around violative content on Twitch cover similar content as the GARM sensitive content categories; however, due to 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. For example, "gore" categories were counted both in "Death, injury, military conflict," and "Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust." Where applicable, we detail in our commentary sections the Twitch enforcement categories mapped to appropriate GARM categories.

We chose not to report on two categories - Arms and Ammunition, and Debated Sensitive Issues. This is the first report where we've included data for Online Piracy which encompasses enforcements against advertisements for piracy, sharing pirated content, and viewership tampering.

Twitch

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. This includes content that does not fall into a GARM sensitive content category, but still violates our guidelines.

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. We only look at content types that are directly indicative of violative content, namely live streams, VODs, clips, and chat. For chat enforcement, we look at a 2 minute window before and after a violative chat is reported to count the violative HWs. This is because chat on Twitch is ephemeral and is expected to disappear from the view quickly.

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:

1. 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).
2. 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 since 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.
3. 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.

Enforcement Metrics

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 are responsible for reviewing these reports and issuing the appropriate enforcements for verified violations. The type of enforcement issued is based on a number of factors and can range from a warning, to a timed suspension, to an indefinite suspension. If there is recorded content associated with the violation, such as a recorded video (VOD) or a clip, that content is removed. That said, most enforcements do not require content removal, given that apart from the report, there is no longer a record of the violation — the live, violative content is already gone. For this reason, we believe the most appropriate measure of our safety efforts is the total number of enforcements issued, and that is how we have oriented the following sections of this report.

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	Previous Period	Commentary
	% of total HW	**% of total HW**	
Adult and Explicit Sexual Content	**0.01%**	**0.01%**	We limit community exposure to content that is not appropriate for all audiences. This includes prohibiting content that involves nudity, and sexually explicit content. These are standards across live, image, and game content. For more information on our policies, please see our Community Guidelines.
Crime and Harmful Acts to Individuals and Society, Human Right Violations	**<0.01%**	**0.01%**	In an effort to limit community exposure to content that may be illegal, upsetting or damaging, we prohibit media and conduct 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 more information on our hate and harassment reports and enforcement, please see our Transparency Report.
Death, Injury or Military Conflict	**<0.01%**	**<0.01%**	In an effort to limit community exposure to content that may be illegal, upsetting or damaging, we prohibit 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.
Hate Speech and Acts of Aggression	**0.01%**	**0.05%**	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. For more information on our policies related to hateful conduct, please see our Community Guidelines.
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic, or repulsive content	**0.02%**	**0.01%**	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	Previous Period	Commentary
	% of total HW	% of total HW	
Online Piracy	<0.01%	N/A	This is the first report where we shared our metrics on Online Piracy and we will continue to share this number moving forward. 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 / ecigarettes / Vaping / Alcohol	<0.01%	<0.01%	We do not permit any activity that may endanger your life or lead to your physical harm. This includes illegal use of drugs and dangerous consumption of alcohol.
Spam or Harmful Content	0.02%	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.
Terrorism	<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.
Other Violations	0.20%	0.12%	For more information, on content that violates the Twitch guidelines, as well as more detailed takedown rates, please see our Community Guidelines and Transparency Report.

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	Previous Period	Commentary
	% of total Impressions	% of total Impressions	
Adult and Explicit Sexual Content	**0.01%**	**0.01%**	We limit community exposure to content that is not appropriate for all audiences. This includes prohibiting content that involves nudity, and sexually explicit content. These are standards across live, image, and game content. For more information on our policies, please see our Community Guidelines.
Crime and Harmful Acts to Individuals and Society, Human Right Violations	**0.11%**	**0.14%**	In an effort to limit community exposure to content that may be illegal, upsetting or damaging, we prohibit media and conduct 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 more information on our hate and harassment reports and enforcement, please see our Transparency Report.
Death, Injury or Military Conflict	**0.01%**	**<0.01%**	In an effort to limit community exposure to content that may be illegal, upsetting or damaging, we prohibit 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.
Hate Speech and Acts of Aggression	**0.27%**	**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. For more information on our policies related to hateful conduct, please see our Community Guidelines.
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic, or repulsive content	**0.06%**	**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	Previous Period	Commentary
	% of total Impressions	% of total Impressions	
Online Piracy	<0.01%	N/A	This is the first report where we shared our metrics on Online Piracy and we will continue to share this number moving forward. 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 / ecigarettes / Vaping / Alcohol	<0.01%	<0.01%	We do not permit any activity that may endanger your life or lead to your physical harm. This includes illegal use of drugs and dangerous consumption of alcohol.
Spam or Harmful Content	0.03%	0.01%	We prohibit disruptive activities such as spamming, because these types of activities violate the integrity of Twitch services, and diminish users' experiences on Twitch.
Terrorism	<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.
Other Violations	0.05%	0.04%	For more information, on content that violates the Twitch guidelines, as well as more detailed takedown rates, please see our Community Guidelines and Transparency Report.

Global Alliance for Responsible Media

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	Previous Period	Commentary
	Enforcement Actions	Enforcement Actions	
Adult & Explicit Sexual Content	35,359	27,920	This includes prohibiting content that involves nudity, and sexually explicit content. These are standards across live, image, and game content. For more information on our policies, please see our Community Guidelines. We continue to invest in the hiring and training of our Law Enforcement Response (LER) team, 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.
Crime and Harmful Acts to Individuals and Society, Human Right Violations	80,406	114,344	In an effort to limit community exposure to content that may be illegal, upsetting or damaging, we prohibit media and conduct 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 more information on our hate and harassment reports and enforcement, please see our Transparency Report. Username-related violations were all classified under the same category. However, with our Usernames Policy update, we specified which policies they violated and recategorized them under the relevant username violation-type (i.e. terrorism related usernames are now terrorism violations).
Death, Injury or Military Conflict	4,006	3,932	In an effort to limit community exposure to content that may be illegal, upsetting or damaging, we prohibit 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.
Hate Speech and Acts of Aggression	115,578	102,682	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. For more information on our policies related to hateful conduct, please see our Community Guidelines.
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic, or repulsive content	31,723	31,661	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.

GARM Global Alliance for Responsible Media

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	Previous Period	Commentary
	Enforcement Actions	**Enforcement Actions**	
Online Piracy	513	N/A	This is the first report where we shared our metrics on Online Piracy and we will continue to share this number moving forward. 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	17	18	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	1,047,949	2,151,932	Our internal teams banned over 13 million disruptive bot accounts which resulted in a massive decrease in enforcement actions for H2 2021. 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. 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.
Terrorism	171	34	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. With the update to our Usernames Policy, usernames that glorify or promote acts of terrorism or terrorists now count as terrorism violations. Due to this change, we anticipated the large increase in enforcements.
Other Violations	502,317	254,148	For more information, on content that violates the Twitch guidelines, as well as more detailed takedown rates, please see our Community Guidelines and Transparency Report. 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.

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**	**Commentary**
Adult & Explicit Sexual Content				
Arms & Ammunition				
Crime and Harmful Acts to Individuals and Society, Human Right Violations				
Death, Injury or Military Conflict				
Online piracy				We are continuing to ramp up our capacity to process appeals. 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.
Hate speech and acts of aggression	6.13%	1.76%	1.76% || 2.68%	
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				

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.

GARM Global Alliance for Responsible Media

GARM Aggregated Measurement Report

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:

Question	Authorized Metric	YouTube	Facebook	Instagram	Twitter	TikTok	Pinterest	Snapchat	Twitch
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
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
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
	Removal of violating accounts by views	Authorized Metric	Authorized Metric	Not Submitted	Next Best Measure	Authorized Metric	Authorized Metric	Authorized Metric	Authorized Metric
	Removal of violating accounts	Authorized Metric	Authorized Metric	Not Submitted	Authorized Metric	Not Submitted	Authorized Metric	Authorized Metric	Authorized Metric
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
	Reinstatements (pieces of content)	Authorized Metric	Authorized Metric	Authorized Metric	Not Submitted	Not Submitted	Authorized Metric	Not Submitted	Authorized Metric

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

	Q1 2021	Q2 2021	Q3 2021	Q4 2021	Q1 2022	Q2 2022
▶ (YouTube)			PREVIOUS PERIOD		LATEST PERIOD	
(Facebook)			PREVIOUS PERIOD		LATEST PERIOD	
(Instagram)			PREVIOUS PERIOD		LATEST PERIOD	
(Twitter)	PREVIOUS PERIOD		LATEST PERIOD			
(TikTok)			PREVIOUS PERIOD		LATEST PERIOD	
(Pinterest)			PREVIOUS PERIOD		LATEST PERIOD	
(Snapchat)	PREVIOUS PERIOD		LATEST PERIOD			
(Twitch)			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
Adult & Explicit Sexual Content	• Nudity & Sexual Content • Child Safety	• Adult Nudity and Sexual Activity, • Child Sexual Exploitation, Abuse and Nudity, • Sexual Solicitation	• Adult Nudity and Sexual Activity, • Child Sexual Exploitation, Abuse and Nudity, • Sexual Solicitation	• Non-Consensual Nudity • Sensitive Media • Child Sexual Exploitation	• Minor safety – sexual exploitation of minors • Adult nudity and sexual activities	• Adult Sexual Services • Adult Content	• Sexually Explicit Content	• Nudity, Pornography, and Other Sexual Content
Arms & Ammunition	• Firearms	• Violence and Incitement • Restricted Goods and Services	• Violence and Incitement • Restricted Goods and Services	• Illegal or certain regulated good or services	• Illegal activities and regulated goods - weapons	• Dangerous Goods and Activities	• Regulated Goods	• Violence and Threats
Crime & Harmful acts to individuals and Society, Human Right Violations	• Harmful or Dangerous Content • Hate Speech • Harassment or cyberbullying	• Adult Nudity and Sexual Activity • Violence and Incitement • Bullying and Harassment • Violent and Graphic Content • Child Sexual Exploitation, Abuse and Nudity • Suicide and Self-Injury • Dangerous Individuals and Organizations • Restricted Goods and Services	• Adult Nudity and Sexual Activity • Violence and Incitement • Bullying and Harassment • Violent and Graphic Content • Child Sexual Exploitation, Abuse and Nudity • Suicide and Self-Injury • Dangerous Individuals and Organizations • Restricted Goods and Services	• Violence • Abuse and harassment	• Illegal activities and regulated goods -criminal activities	• Child Sexual Exploitation • Self-Harm • Harassment & Criticism	• Threatening / Violence / Harm:	• Self-Destructive Behaviour • Hateful Conduct and Harassment
Death, Injury or Military Conflict	• Violent or Graphic Content • Harmful or Dangerous Content • Suicide & Self-Injury	• Violence and Incitement • Violent and Graphic Content • Suicide and Self-Injury	• Violence and Incitement • Violent and Graphic Content • Suicide and Self-Injury	• Promoting Self-harm	• Violent and Graphic Content	• Graphic Violence and Threats	• Threatening / Violence / Harm	• Violence and Threats • Extreme Violence, Gore, and Other Obscene Content
Online piracy	• Fake Engagement • Impersonation • Sale of illegal or regulated goods or services • YouTube Terms of Service	• Intellectual Property Copyright • Intellectual Property Counterfeit • Intellectual Property Trademark	• Intellectual Property Copyright • Intellectual Property Counterfeit • Intellectual Property Trademark	• Copyright • Trademark	• Integrity and authenticity – intellectual property violations	• Copyright • Trademark	• Spam	• Spam, Scams, and Other Malicious Content
Hate speech & acts of aggression	• Hate Speech	• Hate speech • Bullying and Harassment • Dangerous Individuals and Organizations	• Hate speech • Bullying and Harassment • Dangerous Individuals and Organizations	• Hateful Conduct	• Hate Speech • Hateful Behavior	• Hateful Activities	• Threatening / Violence / Harm	• Hateful Conduct and Harassment
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	• Violent or Graphic Content • Age Restriction	• Hate Speech • Bullying and Harassment	• Hate Speech • Bullying and Harassment	• Sensitive Media	• Hateful Behavior – Slurs • Harassment & Bullying	• Harassment & Criticism		• Extreme Violence, Gore, and Other Obscene Content
Illegal drugs, tobacco, e-cigarettes, vaping	• Sale of Illegal or Regulated Goods or Services • Harmful or dangerous content	• Regulated Goods: Drugs	• Regulated Goods: Drugs	• Illegal or certain regulated goods or services	• Illegal activities and regulated goods – drugs, controlled substances, alcohol and tobacco	• Dangerous Goods and Activities	• Regulated Goods	• Self-destructive behaviour
Spam & Malware	• Spam, Deceptive Practices, scams, and misinformation	• Spam	• Spam	• Private information • Impersonation • Platform manipulation	• Integrity and authenticity – spam and fake engagement	• Spam	• Spam	• Spam, Scams, and Other Malicious Content
Terrorism	• Violent criminal organizations	• Dangerous Organizations: Terrorism • Dangerous Organizations: Organized Hate	• Dangerous Organizations: Terrorism • Dangerous Organizations: Organized Hate	• Terrorism or Violent Extremism	• Violent Extremism • Dangerous individuals and organizations - Terrorists and terrorist organizations	• Violent Actors	• Terrorism	• Violence and Threats
Debated Sensitive Social Issues		• Hate Speech • Bullying and Harassment	• Hate Speech • Bullying and Harassment		• Hateful Behavior	• Civic Misinformation • Conspiracy Theories • Medical Misinformation • Climate Misinformation		
Other	• Any categories not specifically accounted for in the above (e.g. multiple policy violations)	• COVID-19 and Vaccine Policy and Protections	• COVID-19 and Vaccine Policy and Protections	• Covid Integrity • Covid-19 Misleading Information				• Suspension Evasion • Unauthorized Sharing of Private Information • Impersonation • Cheating in Online Games

WFA