

GARM Aggregated Measurement Report



Volume 3 | May 2022

Contents

- 03** Aggregated Measurement Report
- 04** Using the GARM Aggregated Measurement Report
- 06** Executive Summary
- 08** YouTube
- 18** Facebook
- 35** Instagram
- 47** Twitter
- 56** TikTok
- 68** Pinterest
- 78** Snapchat
- 88** Twitch
- 97** Appendices & FAQ

Creating the GARM Aggregated Measurement Report

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

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

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

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

Why are we creating this report?

YouTube, Facebook, and Twitter all provided content policy reporting in 2018. Over time more digital media platforms have adopted this practice with the goal of communicating effective content moderation practices to several stakeholder audiences, ranging from regulators to NGOs to advertisers. With GARM's focus on societal safety and media industry

sustainability, we want to more accurately communicate progress and challenges in individual and collective work to eliminate harmful content from ad supported media. We've created the GARM Aggregated Measurement Report with advertising industry stakeholders in mind, and are delivering value through the following 5 steps:

Creating a single access point

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

Establishing a framework for industry focus

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

Defining a set of quality metrics to answer critical questions

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

Creating a link between policy to established categories

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

Providing contextual insights on data

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

Using the GARM Aggregated Measurement Report

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

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

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

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

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

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

What is the framework for the report?

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

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

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

CORE QUESTION	AUTHORIZED METRIC	DEFINITION + OVERVIEW	RATIONALE
How safe is the platform for consumers?	Prevalence of violating content or Violative View Rate	The percentage of views that contain content that is deemed as violative	Establishes a ratio based on typical user content consumption. Prevalence or Violative View Rate examines views of unsafe/violating content as a proportion of all views.
How safe is the platform for advertisers?	Prevalence of violating content or Advertising Safety Error Rate	The percentage of views that contain content that is deemed as violative The percentage of views of monetized content that contain violative content	Monetization prevalence examines unsafe content viewed as a proportion of monetized content viewed
How effective is the platform in policy enforcement?	Removals of Violating Content + Removal of Violating Accounts Removals of Violating Content expressed by how many times it has been viewed	Pieces of violating content removed Accounts removed due to repeat policy violation Pieces of violating content removed categorized by how many times they were viewed by users	Platform teams spend a considerable amount of time removing violating content and bad actors from their platforms – the magnitude of the efforts should be reported to marketers. It is also important to 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 Oversight Working Group to evolve the report for approval by the GARM Steer Team.

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

Executive Summary

It's been a year since we launched the first Aggregated Measurement Report, and we are happy to release volume three, which features data from participating platforms from Q3 2020 through Q4 2021. This new installment features the following key advancements for the report and underlying reporting efforts:

01

First, YouTube this week becomes the only media platform to receive independent accreditation for metrics intended to report on the safety of monetized content. This was done as part of the MRC Content Level Brand Safety Controls Audit and in YouTube's annual recertification for May 2022. YouTube is the only platform at present to receive the accreditation and with monetization metrics as part of the audit, which helps verify the transparency reporting done by a platform relative to the content safety of their advertising business

02

Second, we see that platforms like Facebook and Instagram continue to publish key measures at additional content categories. This is good progress and helps industry participants understand the effects of safety policy enforcement better

03

Third, we are seeing more and more platforms report on actions specific to misinformation, with many of them detailing specific focus areas (e.g., Medical Misinformation/COVID). This is a good precedent for where we envision the industry and GARM's work going forward – detail on focused enforcement areas

04

Lastly, we've included a 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

Finally, it is important to note that this report does not include data pertaining to the War on Ukraine. The data relative to platform enforcement on content moderation relative to the War on Ukraine will first feature in individual platform ongoing reports and will then feature later in the next volume of GARM Aggregated Measurement Report in November 2022.

LEARNING 1: Enforcement in GARM Categories Spam & Malware and Adult & Explicit Sexual Content continues to be the largest and most automated. All the platforms that participated in this volume of the GARM Aggregated Measurement Report indicated that either Spam or Adult Content were the leading enforcement areas in terms of content blocked, removed, and actors actioned. This trend for these two categories was seen in 7-out-of-7 of the platforms in our report. Spam & Malware represent a much larger category when comparing the two. Additionally, given improvements in technology features like machine learning which allow for scaled enforcement in these areas, we are also seeing advanced enforcement in these two content categories. For example, when looking at platforms that share relevant data pertaining to adult and explicit sexual content, we see that Pinterest reports a hybrid deactivation rate of 99% and Facebook a proactive rate of 98%.

LEARNING 2: Highly nuanced content such as Crime & Harmful acts to Individuals and Society are highly reliant on context and remain the most manual. This sits in contrast with our learning on Spam and Explicit Sexual Content, which is highly automated. We see proactive block rates for these types of instances to be low and manual deactivations to be high relative to other content types – for instance Facebook reports a proactive rate of 59% while Pinterest reports that 62% of Pins deactivated for graphic violence and threats are actioned manually. These types of harmful content are highly contextual in nature and require human moderation beyond technology, and in many cases user reporting. As noted in our last Aggregated Measurement Report, some platforms have introduced disruptive techniques known as friction to discourage this type of content from posters. We see pronounced enforcement metrics pertaining to removals for these content categories in these platforms. Individualized safety requires more resourcing and should be an area for future resourcing and innovation.

Executive Summary

LEARNING 3: Content areas like Misinformation and Self-Harm are coming into focus in reporting moderation efforts. Harmful content categories like Self Harm are seeing upticks in enforcement data, as seen in Snap taking a move towards dedicating it as a category to start quantifying the risk and remediation. We see a similar trend in Misinformation. Platforms approaches towards misinformation have scaled up into increasingly more structured and complex operations. Recent accelerants to existing work have been external events in the public health and elections area, increased digital media regulation, and aided by work being done within GARM. As the platforms have continued in this area, we are seeing more consistent tracking and sharing of this data. For instance, we are seeing structured sharing of enforcement of this – 6 of the 7 platforms are all now sharing data on enforcement on Misinformation. Of note, Pinterest and Twitter break out reporting on flags and removals for specific Misinformation sub-policies – Medical and Civic Misinformation. Other platforms report on Misinformation enforcement as one holistic category or as nested in other policy areas, e.g. Harmful and Dangerous Content. This depends on how each platforms' policies are developed and enforced. We should expect more structured reporting on this category in future volumes of the GARM Aggregated Measurement Report, as we introduce Misinformation as a Harmful Content Category in June.

A look forward to Volume 4 and beyond:

Additional Category Inclusion: In the coming weeks, GARM will announce an update to the harmful content categories and expand the GARM Brand Safety Floor and Suitability Framework to include Misinformation. As we've seen platforms report on this in an ad-hoc fashion, we are anticipating that participants will include this in a more structured format over time, in part supported by regulatory advances by governments in London and Brussels.

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 this accreditation. We eagerly await more platforms taking this step.

Placement Data: We recently expanded GARM membership to adtech companies, many of whom help advertisers, agencies, and platforms in the placement of ads. We will embark on a journey to incorporate data from these providers in the future to help answer the critical questions around safe monetization.

YouTube

Our Commitment to Responsibility

At YouTube, we work hard to maintain a safe and vibrant community. **Responsibility remains our #1 priority, and we approach this work from several angles: we remove violative content, raise up authoritative voices, reduce the spread of content that brushes right up against our policy line and reward trusted, eligible creators and artists.**

YouTube has 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 misinformation that presents serious risk of egregious harm. 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 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 investments in machine learning, **in H2 2021 we were able to detect >93% of all violative content on YouTube by automated flagging – with more than two-thirds of flagged content removed with fewer than 10 views.**

Content flagged by users is 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.

We have also made huge investments in other areas of critical importance, like transparency. Several years ago, **YouTube became the first major platform to launch a transparency report and offer insights on these removals, including the number of videos removed for policy violations, how that violative content was first identified, reasons for removal, and more.** This includes the publication of our **Violative View Rate Metric, VVR**, which helps us determine what percentage of views on YouTube comes from content that violates our Community Guideline Policies. VVR is included in this report as a GARM Authorized metric.

Every quarter, our Transparency Report showcases data that demonstrates the vast impact of our enforcement work and the progress we've made. We've pulled critical insights from our last four transparency reports into this resource. This includes:

- Flagging (human and automated)
- Video, channel, and comment removals
- Appeals and reinstatements
- 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. **For the purposes of the GARM Aggregated Measurement Report Volume 3, we have included quarterly data from full year 2021, represented as bi-annual aggregations (H2 2021 & H1 2021).**

In addition to our Community Guidelines, we enforce a complementary 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 MRC accreditation for content-level brand safety, we also publish our effectiveness at enforcing Ad Friendly Guidelines [here](#).

In May 2022, the Media Rating Council (MRC) recertified YouTube's accreditation for content-level Brand Safety, certifying that YouTube in-stream video ads and the Advertiser Safety Error Rate metric adhere to the industry standards for content level brand safety processes and controls.

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

Report Insights

In this report, YouTube is **proud to answer all four core questions using GARM Authorized metrics**, as we have in previous GARM Aggregated Measurement Report volumes.

July through December 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. YouTube terminated over 8.6 million channels for violating our Community Guidelines and the overwhelming majority 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. In September 2021, Professor Arnold Barnett at MIT Sloan published a report to assess the completeness and the appropriateness of our VVR metric. Barnett found the methodology for VVR statistically sound and saw VVR as an accurate measure of our enforcement of Community Guidelines.

Lastly, YouTube removed more than 2 billion comments, the majority of which were spam; 99% of removed comments were detected automatically. Over 457k video removals were appealed, and we reinstated <133k of those videos.

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 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.

January through June 2021

Between January and June 2021, YouTube removed over 15 million videos for violating Community Guidelines. The vast majority (>94%) of these videos were first flagged by machines rather than humans. YouTube terminated over 6.4 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 0.16-0.18% in Q1 2021 to 0.19-0.21% in Q2 2021. This means that out of every 10,000 views on YouTube in Q2, only 19-21 came from violative content in Q2.

Lastly, YouTube removed more than 2 billion comments, the majority of which were spam; 99% of removed comments were detected automatically. Just over 434k video removals were appealed, and we reinstated ~118k of those videos.

In February 2021, the Media Rating Council (MRC) **granted the digital industry's first content level Brand Safety Accreditation to YouTube**. The Media Rating Council accreditation states that YouTube in-stream video ads adhere to the industry standards for content level brand safety processes and controls. 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

Methodology for Metrics

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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



Question 1: How safe is the platform for consumers?

Authorized Metric: Violative View Rate

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

GARM Metric	Latest Period		Previous Period	
	Q4 2021	Q3 2021	Q2 2021	Q1 2021
Violative View Rate	0.12-0.14%	0.09-0.11%	0.19-0.21%	0.16-0.18%

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



Question 2: How safe is the platform for advertisers?

Authorized Metric: Advertising Safety Error Rate



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

GARM Metric	Latest Period		Previous Period	
	Q4 2021	Q3 2021	Q2 2021	Q1 2021
Advertising Safety Error Rate	<1%	<1%	<1%	<1%



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

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

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

YouTube Community Guidelines

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

YouTube Policy	Content Actioned ¹		Actors Actioned ²		Comments Actioned	
	Latest Period Q3 & Q4 2021	Previous Period Q1 & Q2 2021	Latest Period Q3 & Q4 2021	Previous Period Q1 & Q2 2021	Latest Period Q3 & Q4 2021	Previous Period Q1 & Q2 2021
Nudity or sexual	1,839,103	2,986,734	313,035	316,217	4,908,817	2,962,591
Child safety	3,168,476	7,006,199	95,007	72,904	299,902,525	546,754,050
Harmful or dangerous	585,755	511,374	36,201	16,927	17,713	378,685
Promotion of violence and violent extremism	322,751	513,908	21,803	25,587	291,379	234,179
Harassment and cyberbullying	606,582	383,781	129,778	164,679	292,998,647	334,456,305
Violent or graphic	2,202,748	2,547,113	27,916	1,799	320,893	182,725
Spam, deceptive practices, scams and misinformation	947,384	1,644,919	7,892,531	5,675,633	1,669,336,620	1,210,489,698
Hateful or abusive	203,532	173,190	78,844	75,844	94,132,595	99,063,773
Impersonation	n/a	n/a	51,791	52,290	n/a	n/a
Other	107,766	81,194	9,411	16,030	31,123	6

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	Previous Period
	Q3 & Q4 2021	Q1 & Q2 2021
Total Video Removals	9,984,097	15,848,412
Removed videos by views: 0 views	34.8%	31.8%
Removed videos by views: 1-10	36.6%	38.1%
Removed videos by views: 10+	28.6%	30.1%



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

Authorized Metric: Appeals, Reinstatements

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

GARM Metric	Latest Period	Previous Period
	Q3 & Q4 2021	Q1 & Q2 2021
Content Appealed: Videos	457,426	434,022
Content Reinstated: Video	132,737	118,716



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 <ul style="list-style-type: none"> Defines content that can monetize. Aligned with YouTube's Ad Friendly Guidelines, a higher bar than Community Guidelines. 	Relevant YouTube Community Guidelines <ul style="list-style-type: none"> Governs content that can live on YouTube. Our Community Guidelines Enforcement Report measures our enforcement of these guidelines.
Adult & Explicit Sexual Content <ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> 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 groups or physical traits.
Death, Injury or Military Conflict <ul style="list-style-type: none"> 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.



Mapping GARM Categories and Monetization to YouTube Community Policy-level Reporting (continued)

<p>Online piracy</p> <ul style="list-style-type: none"> • Pirating, Copyright infringement, & Counterfeiting 	<p>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.</p> <p>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.</p> <p>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</p> <p>YouTube's Terms of Service Also covered in YouTube's Terms of Service</p>
<p>Hate speech & acts of aggression</p> <ul style="list-style-type: none"> • 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. 	<p>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</p>
<p>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</p> <ul style="list-style-type: none"> • Excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult. 	<p>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.</p> <p>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.</p>
<p>Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol</p> <ul style="list-style-type: none"> • 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 	<p>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</p> <p>Harmful or dangerous content YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death.</p>
<p>Spam or Harmful Content</p> <ul style="list-style-type: none"> • Malware/Phishing 	<p>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.</p>
<p>Terrorism</p> <ul style="list-style-type: none"> • Promotion and advocacy of graphic terrorist activity involving defamation, physical and/or emotional harm of individuals, communities, and society 	<p>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.</p>
<p>Debated Sensitive Social Issue</p> <ul style="list-style-type: none"> • Insensitive, irresponsible and harmful treatment of debated social issues and related acts that demean a particular group or incite great conflict 	
<p>Other</p>	<p>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.</p>



Our Commitment to Responsibility and Transparency on Facebook and Instagram

We want Facebook and Instagram to be places where people have a voice. To create conditions where everyone feels comfortable expressing themselves, 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. These policies either meet or, in many cases, exceed the [GARM Brand Safety Floor](#). We don’t allow anything 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. 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 can 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.

In March, we also updated the [Misinformation](#) section to our Community Standards, by consolidating different aspects of our misinformation policy into one easy-to-find location. While none of our substantive policies have changed; this new section is about making our policies easier to access and understand in context. This is in part a response to the Oversight Board’s request that we do so.

Since the beginning of the pandemic, we displayed warnings on more than 195 million pieces of COVID-related content on Facebook that our fact-checking partners rated. And, we’ve removed more than 25 million pieces of content from Facebook and Instagram globally for violating our [policies on COVID-19-related misinformation](#). We’ve also taken action against people who repeatedly post content that breaks our rules. Since the beginning of the pandemic, we have removed over 3,000 accounts, pages, and groups for repeatedly violating our rules against spreading COVID-19 and vaccine misinformation.

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](#).

General trends for H1 2021 and H2 2021 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, the content actioned (and the percentage of that we found proactively), and the 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 hate speech on Facebook has continued to decrease since we first began reporting it, in Q4 2021 it was 0.02%-0.03% or 2-3 views of content per 10,000 views down from 0.10%-0.11%. This is due to improvements in proactively detecting hate speech and ranking changes in Feed.
- Hate speech content removal has increased over 15X on Facebook and Instagram since 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 nine out of 12 on Instagram.
- We now include 14 policy areas on Facebook and 12 on Instagram, and have added new metrics on appeals, restores, and five new prevalence metrics (Violence and Incitement on Facebook and Instagram; Bullying and Harassment on Facebook and Instagram; and Hate Speech on Instagram) and introduced Violence and Incitement as a new reporting area with all metrics on Facebook and Instagram.



In [March](#), we introduced the [Family Center](#), a new place for parents and guardians to access supervision tools and resources from leading experts. [Supervision tools](#) are available on Instagram today and will begin rolling out in VR in May. This is the first step in a longer term journey to develop intuitive supervision tools, informed by experts, teens and parents. Our vision for the Family Center is to eventually allow parents and guardians to help their teens manage experiences across Meta technologies, all from one central place.

Bullying and harassment

In our Q3 2021 report, we included prevalence for bullying and harassment on Facebook and Instagram for the first time. We work hard to enforce against this content while also equipping our community with tools to protect themselves in ways that work best for them.

When it comes to [bullying and harassment](#), context and intent matter. We have developed AI systems that can identify many types of bullying and harassment across our platforms. However, bullying and harassment is a unique issue area because determining harm often requires context, including reports from those who may experience this behavior. Bullying and harassment are often very personal — it shows up in different ways for different people. What can be a light-hearted comment between friends, can be bullying or harassment in another context. As a result, detecting such bullying can be more challenging than other types of violations.

While we are always working to improve our technology, our metrics (particularly those for proactive rate and prevalence) reflect the reality of having to rely on reports from our community. Prevalence metrics allow us to track, both internally and externally, how much violating content people are seeing on our apps. Prevalence, in turn, helps us determine the right approaches to driving that metric down, whether it's through updating our policies, products or tools for our community.

One recent tool we've deployed on both Facebook and Instagram is adding warning screens to educate and discourage people from posting or commenting in ways that could be bullying and harassment. On [Instagram](#), about 50% of the time the comment was edited or deleted by the user based on these warnings. We've made investments in our bullying and harassment tools on Instagram, working to reduce the number of times this type of content is seen particularly by our younger user base. We have a number of resources where you can learn more about the bullying and harassment prevention work we do on our platform, including our [Bullying Prevention Hub](#).

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 2021 (our latest report) on Facebook and Instagram

Prevalence of harmful content on Facebook and Instagram remained relatively consistent or decreased from Q3 to Q4 across most of our policy areas, meaning the vast majority of content that users encounter does not violate our standards. Hate speech prevalence on Facebook was 0.02% - 0.03% in Q4 2021, down from 0.10% - 0.11% when we began reporting it. Hate speech prevalence on Instagram was reported for the second time and stayed relatively consistent; in Q3 2021 it was 0.02%, and in Q4 2021 it was 0.02% - 0.03%.

Here are a few examples of the ways our holistic approach to creating safer spaces impacted the results in the Q4 2021 report:

- **Advances in AI's impact on hate speech and suicide and self-injury:** Advancements in AI technologies are central to our ability to continue to correctly take action on violating content more quickly. Our continued progress in the reduction of hate speech can be directly attributed to the [Few Shot Learner](#) tool. Due to the expansion of our media-matching technology to identify and remove old, violating content, we were also able to increase content actioned on suicide and self-injury on Instagram.
- **People's impact on restored content:** It's important we're transparent with people about when content is removed, what policies it's violating and to make it simple for them to appeal, as we don't always get it right. Following the launch of [Account Status](#) on Instagram, we saw an increase of restored content across several policy areas, including violence and incitement, hate speech, bullying and harassment and adult nudity and sexuality.
- **Product design's impact on creating safe spaces:** Our product design teams play an important role in our effort to reduce harmful content. By carefully designing our social media products, we can promote greater safety while providing people with context, control and the room to share their voice. For example, when applied, informative overlays and labels provide more context and limit exposure to misinformation or potentially sensitive (e.g. graphic) content.



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](#).



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 a metric externally we goes 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, with plans to test in the second half of the 2022 before rolling out more broadly in early 2023. These controls will be subject to an MRC audit once they have been 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 will work together to develop a 3rd party post campaign solution to measure and verify the suitability of adjacent content to ads in Feed, with the goal of starting with small scale testing in the third quarter of this year and moving to limited availability in the fourth quarter of 2022.

Independent Verification

Community Standards Enforcement Report audit

We believe that no company should grade its own homework. To validate that our metrics are measured and reported correctly, we are conducting an audit that will cover the Q4 2021 Community Standards Enforcement Report, and anticipate releasing the results in May. The assessment will be completed by EY, and we encourage all or our peers to undergo similar assessments.

MRC audit

We are committed to independent verification and support the MRC as the auditor of our monetized solutions in advertising. We are currently in process with the MRC on the audit of our Content Monetization Policies and Brand Safety & Suitability Controls (for Facebook In Stream and Instant Article placements, not yet including Facebook and Instagram Feed). We expect to complete additional efforts related to the audit of our Content Monetization Policies and Brand Safety & Suitability Controls for MRC review in Q2 of 2022.

As outlined in the 2022 roadmap, the first party suitability controls we are currently building for Facebook and Instagram Feed are being developed with an understanding that the efficacy of the control once generally available will be verified through an MRC audit.

TAG certification

We achieved our Brand Safety Certified Seal from TAG in 2020 and have extended this certification in 2021 - we are currently closing our evaluation with the auditor to continue the Brand Safety Certified seal across Facebook, Instagram, and Audience Network.



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

Other Facebook Policies Floor does not address

- [Privacy Violations](#)
- [Account Integrity and Authentic Integrity](#)
- [Inauthentic Behavior](#)
- [Misinformation](#)
- [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.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2021	Q3 2021	Q2 2021	Q1 2021	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	0.03%	0.02%-0.03%	0.04%	0.03%-0.04%	<p>Adult Nudity and Sexual Activity: Prevalence has remained relatively consistent.</p> <p>In 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. We cannot estimate prevalence for these right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.</p>
	Child Endangerment: Nudity and Physical Abuse	N/A	N/A	N/A	N/A	
	Child Endangerment: Sexual Exploitation	N/A	N/A	N/A	N/A	
Arms & Ammunition	Regulated Goods: Firearms	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	<p>Violence & Incitement: Prevalence was 0.03% to 0.04% in Q4 2021, which marks a decrease from Q3 2021 due to a decrease in overall and violating comments. We first reported prevalence in Q3 2021.</p>
	Violence & Incitement	0.03%-0.04%	0.04%-0.05%	N/A	N/A	



Question 1: How safe is the platform for consumers?

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2021	Q3 2021	Q2 2021	Q1 2021	
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	0.03%	0.02%-0.03%	0.04%	0.03%-0.04%	<p>Violence & Incitement: Prevalence was 0.03% to 0.04% in Q4 2021, which marks a decrease from Q3 2021 due to a decrease in overall and violating comments. We first reported prevalence in Q3 2021.</p> <p>Violent and Graphic Content: Prevalence was between 0.03-0.04% of views in Q1 2021, which marks a decrease from Q4. This was due to ranking changes to personalize content for users and reduce problematic content in Feed.</p> <p>Bullying and Harassment: Prevalence was 0.11% to 0.12% in Q4 2021, which marks a decrease from Q3 2021 due to refinements we made to our policies to identify this content. We first reported prevalence in Q3 2021.</p> <p>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. We cannot estimate prevalence for these right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.</p>
	Violence & Incitement	0.03%-0.04%	0.04%-0.05%	N/A	N/A	
	Violent and Graphic Content	0.03%-0.04%	0.04%	0.03%-0.04%	0.03%-0.04%	
	Bullying and Harassment	0.11%-0.12%	0.14%-0.15%	N/A	N/A	
	Child Nudity and Sexual Exploitation	N/A	N/A	N/A	Less than 0.05%	
	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 0.06%	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
		Q4 2021	Q3 2021	Q2 2021	Q1 2021	
Death, Injury or Military Conflict	Violent and Graphic Content	0.03%-0.04%	0.04%	0.03%-0.04%	0.03%-0.04%	Violent and Graphic Content: Prevalence was between 0.03-0.04% of views in Q1 2021, which marks a decrease from Q4. This was due to ranking changes to personalize content for users and reduce problematic content in Feed.
	Violence and Incitement	0.03%-0.04%	0.04%-0.05%	N/A	N/A	
	Suicide and Self Injury	Less than 0.05%	Less than 0.05%	Less than 0.06%	Less than 0.05%	
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	We do not report prevalence of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	
Hate speech & acts of aggression	Hate Speech	0.02%-0.03%	0.03%	0.05%	0.05%-0.06%	Hate Speech: Prevalence of hate speech content was about 0.03% in Q3 2021, which marks a decrease from Q2 2021 due to our proactive detection technology and ranking changes to personalize content for users and reduce problematic content in Feed. Prevalence of hate speech content was 0.05% in Q2 2021, which marks a decrease from Q1 2021 due to proactive detection of comments and ranking changes to personalize content for users and reduce problematic content in Feed.
	Dangerous Organizations: Terrorism	Less than 0.05%	Less than 0.05%	Less than 0.07%	Less than 0.06%	
	Dangerous Organizations: Organized Hate	N/A	N/A	N/A	N/A	Bullying and Harassment: Prevalence was 0.11% to 0.12% in Q4 2021, which marks a decrease from Q3 2021 due to refinements we made to our policies to identify this content. We first reported prevalence in Q3 2021.
	Bullying and Harassment	0.11%-0.12%	0.14%-0.15%	N/A	N/A	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.



Question 1: How safe is the platform for consumers?

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2021	Q3 2021	Q2 2021	Q1 2021	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	0.02%-0.03%	0.03%	0.05%	0.05%-0.06%	Hate Speech: Prevalence of hate speech content was about 0.03% in Q3 2021, which marks a decrease from Q2 2021 due to our proactive detection technology and ranking changes to personalize content for users and reduce problematic content in Feed. Prevalence of hate speech content was 0.05% in Q2 2021, which marks a decrease from Q1 2021 due to proactive detection of comments and ranking changes to personalize content for users and reduce problematic content in Feed.
	Bullying and Harassment	0.11%-0.12%	0.14%-0.15%	N/A	N/A	Bullying and Harassment: Prevalence was 0.11% to 0.12% in Q4 2021, which marks a decrease from Q3 2021 due to refinements we made to our policies to identify this content. We first reported prevalence in Q3 2021.
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 0.07%	Less than 0.06%	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%-0.03%	0.03%	0.05%	0.05%-0.06%	Hate Speech: Prevalence of hate speech content was about 0.03% in Q3 2021, which marks a decrease from Q2 2021 due to our proactive detection technology and ranking changes to personalize content for users and reduce problematic content in Feed. Prevalence of hate speech content was 0.05% in Q2 2021, which marks a decrease from Q1 2021 due to proactive detection of comments and ranking changes to personalize content for users and reduce problematic content in Feed.
	Bullying and Harassment	0.11%-0.12%	0.14%-0.15%	N/A	N/A	Bullying and Harassment: Prevalence was 0.11% to 0.12% in Q4 2021, which marks a decrease from Q3 2021 due to refinements we made to our policies to identify this content. We first reported prevalence in Q3 2021.



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

Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Facebook

GARM Category	Relevant Policy	Latest Period				Previous Period			
		Q4 2021		Q3 2021		Q2 2021		Q1 2021	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	27.3m	97.7%	34.7m	98.8%	32.8m	98.9%	31.9m	98.6%
	Child Endangerment: Nudity and Physical Abuse	1.8m	97.5%	1.8m	97.1%	2.3m	97.9%	N/A	N/A
	Child Endangerment: Sexual Exploitation	19.8m	99.0%	21.2m	99.1%	25.6m	99.5%	N/A	N/A
Arms & Ammunition	Regulated Goods: Firearms	1.5m	92%	1.1m	94.1%	1.5m	94%	1.9m	94.8%
	Violence & Incitement	12.4m	96.6%	13.6m	96.7%	N/A	N/A	N/A	N/A
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	27.3m	97.7%	34.7m	98.8%	32.8m	98.9%	31.9m	98.6%
	Violence & Incitement	12.4m	96.6%	13.6m	96.7%	N/A	N/A	N/A	N/A
	Violent and Graphic Content	25.2m	99.5%	26.6m	99.4%	30.1m	99.6%	34.1m	99.6%
	Bullying and Harassment	8.2m	58.8%	9.2m	59.4%	7.9m	54.1%	8.8m	54.1%
	Child Nudity and Sexual Exploitation	N/A	N/A	N/A	N/A	N/A	N/A	25.7m	99.5%
	Child Endangerment: Nudity and Physical Abuse	1.8m	97.5%	1.8m	97.1%	2.3m	97.9%	N/A	N/A
	Child Endangerment: Sexual Exploitation	19.8m	99.0%	21.2m	99.1%	25.6m	99.5%	N/A	N/A
	Suicide and Self-Injury	6.1m	98.8%	8.5m	99%	16.8m	99.3%	5.2m	97.7%
Death, Injury or Military Conflict	Regulated Goods: Firearms	1.5m	92%	1.1m	94.1%	1.5m	94%	1.9m	94.8%
	Violent and Graphic Content	25.2m	99.5%	26.6m	99.4%	30.1m	99.6%	34.1m	99.6%
	Violence and Incitement	12.4m	96.6%	13.6m	96.7%	N/A	N/A	N/A	N/A
Online piracy	Suicide and Self Injury	6.1m	98.8%	8.5m	99%	16.8m	99.3%	5.2m	97.7%
	Intellectual Property: Copyright	1.3m	75%	1.4m	82%	1.58m	83.9%	1.4m	85.3%
	Intellectual Property: Counterfeit	740.7k	76%	760.5k	85%	721.4k	79.6%	506.9k	84.9%
	Intellectual Property: Trademark	189.7k	52%	219.8k	65%	178.9k	70.2%	165.8k	67.3%

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



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

Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Facebook

GARM Category	Relevant Policy	Latest Period				Previous Period			
		Q4 2021		Q3 2021		Q2 2021		Q1 2021	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
Hate speech & acts of aggression	Hate Speech	17.4m	95.9%	22.3m	96.5%	31.5m	97.6%	25.2m	96.7%
	Dangerous Organizations: Terrorism	7.7m	97.7%	10.6m	97.9%	7.1m	99.7%	9m	99.6%
	Dangerous Organizations: Organized Hate	1.6m	96.1%	2m	96.4%	6.2m	97.8%	9.8m	98.6%
	Bullying and Harassment	8.2m	58.8%	9.2m	59.4%	7.9m	54.1%	8.8m	54.1%
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	17.4m	95.9%	22.3m	96.5%	31.5m	97.6%	25.2m	96.7%
	Bullying and Harassment	8.2m	58.8%	9.2m	59.4%	7.9m	54.1%	8.8m	54.1%
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	4m	97.9%	2.7m	96.7%	2.3m	93.1%	3.2m	94.1%
Spam or Harmful Content	Spam	1.2b	99.6%	777.2m	99.6%	794.4m	99.7%	918.9m	99.8%
Terrorism	Dangerous Organizations: Terrorism	7.7m	97.7%	10.6m	97.9%	7.1m	99.7%	9m	99.6%
	Dangerous Organizations: Organized Hate	1.6m	96.1%	2m	96.4%	6.2m	97.8%	9.8m	98.6%
Debated Sensitive Social Issue	Hate Speech	17.4m	95.9%	22.3m	96.5%	31.5m	97.6%	25.2m	96.7%
	Bullying and Harassment	8.2m	58.8%	9.2m	59.4%	7.9m	54.1%	8.8m	54.1%



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
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	Adult Nudity and Sexual Activity: In Q1, we adjusted our media-matching technology and were able to take action on old, violating content.
	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. In Q2 2021 we saw increased content actioned as we improved our proactive detection technology on videos and expanded our media-matching technology on Facebook allowing us to remove more old, violating content, both of which led to the majority of content actioned. Content actioned was published for the first time in our Q2 report, and was 25.7 million pieces of content. Our previous metric for this category considered only a subset of enforcements. With this release, we now include additional categories of content we remove under this policy. Due to these changes, it may not be accurate to compare the metric directly with previous quarters. In Q2 2021, we improved our proactive detection technology on videos and expanded our media-matching technology on Facebook allowing us to remove more old, violating content, both of which led to the majority of content actioned. The proactive rate was 99.5% in Q2 2021.
	Child Endangerment: Sexual Exploitation	Child Endangerment: Sexual Exploitation: Content actioned for child sexual exploitation decreased from 25.6 million pieces of content in Q2 2021 to 20.9 million in Q3 2021. In Q2 2021 we saw increased content actioned as we improved our proactive detection technology on videos and expanded our media-matching technology on Facebook allowing us to remove more old, violating content, both of which led to the majority of content actioned. Content actioned for child nudity and physical abuse was published for the first time in our Q2 report, and was 2.3 million pieces of content. Our previous metric for this category considered only a subset of enforcements. With this release, we now include additional categories of content we remove under this policy. Due to these changes, it may not be accurate to compare the metric directly with previous quarters. The proactive rate was 97.9% in Q2 2021.
Arms & Ammunition	Regulated Goods: Firearms	Regulated Goods: Firearms: 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	
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	Adult Nudity and Sexual Activity: In Q1, we adjusted our media-matching technology and were able to take action on old, violating content.
	Violence & Incitement	
	Violent and Graphic Content	Violent and Graphic Content: Restored content decreased from 8.6K in Q3 2021 to 6.2K in Q4 2021 due to a decrease in restores without appeals on video 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: Content actioned increased from 7.9 million in Q2 2021 to 9.2 million in Q3 2021 as we expanded our proactive detection technology to more languages.
	Child Nudity and Sexual Exploitation	Child Nudity and Sexual Exploitation: From Q2 2021 we've added more data and 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	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. In Q2 2021 we saw increased content actioned as we improved our proactive detection technology on videos and expanded our media-matching technology on Facebook allowing us to remove more old, violating content, both of which led to the majority of content actioned. Content actioned was published for the first time in our Q2 report, and was 25.7 million pieces of content. Our previous metric for this category considered only a subset of enforcements. With this release, we now include additional categories of content we remove under this policy. Due to these changes, it may not be accurate to compare the metric directly with previous quarters. In Q2 2021, we improved our proactive detection technology on videos and expanded our media-matching technology on Facebook allowing us to remove more old, violating content, both of which led to the majority of content actioned. The proactive rate was 99.5% in Q2 2021.
	Child Endangerment: Sexual Exploitation	Child Endangerment: Sexual Exploitation: Content actioned for child sexual exploitation decreased from 25.6 million pieces of content in Q2 2021 to 20.9 million in Q3 2021. In Q2 2021 we saw increased content actioned as we improved our proactive detection technology on videos and expanded our media-matching technology on Facebook allowing us to remove more old, violating content, both of which led to the majority of content actioned. Content actioned for child nudity and physical abuse was published for the first time in our Q2 report, and was 2.3 million pieces of content. Our previous metric for this category considered only a subset of enforcements. With this release, we now include additional categories of content we remove under this policy. Due to these changes, it may not be accurate to compare the metric directly with previous quarters. The proactive rate was 97.9% in Q2 2021.
	Suicide and Self-Injury	Suicide and Self Injury: 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.



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
Death, Injury or Military Conflict	Violent and Graphic Content	Violent and Graphic Content: Restored content decreased from 8.6K in Q3 2021 to 6.2K in Q4 2021 due to a decrease in restores without appeals on video 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	
	Suicide and Self Injury	Suicide and Self Injury: 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.
Online piracy	Intellectual Property: Copyright	We report this metric monthly in a 6 month report 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.
	Intellectual Property: Counterfeit	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: 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 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. Content actioned for terrorism increased from 7.1 million pieces of content in Q2 2021 to 9.8 million in Q3 2021, primarily due to 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 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. Content actioned for organized hate decreased from 6.2 million in Q2 2021 to 2 million in Q3 2021 as we updated our media matching technology to improve the precision of our decisions.
	Bullying and Harassment	Bullying and Harassment: Content actioned increased from 7.9 million in Q2 2021 to 9.2 million in Q3 2021 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	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: Content actioned increased from 7.9 million in Q2 2021 to 9.2 million in Q3 2021 as we expanded our proactive detection technology to more languages.
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 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 9 million pieces of content in Q1 2021 to 7.1 million in Q2 2021, primarily due to an update to our proactive detection technology. This marks a return to pre-Q1 levels. Content actioned increased in Q2 primarily driven by expanding our proactive detection technology to help us detect and review more potential violations, often before anyone sees the content.
	Dangerous Organizations: Organized Hate	Dangerous Organizations: Organized Hate: Content actioned decreased from 9.8 million pieces of content in Q1 2021 to 6.2 million in Q2 2021, primarily due to an update to our proactive detection technology. This marks a return to pre-Q1 levels. Content actioned increased from 6.4 million pieces of content in Q4 2020 to 9.8 million in Q1 2021. In Q1, we adjusted our media-matching technology and were able to take action on old, violating content.
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: Content actioned increased from 7.9 million in Q2 2021 to 9.2 million in Q3 2021 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

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2021			Q3 2021			Q2 2021			Q1 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	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	311.4k	36.7k	262k	272.9k	32.9k	233.6k	536.4k	55.9k	311.9k	390k	73.1k	559.5k	<p>Adult Nudity and Sexual Activity: Appealed content has continued to increase from 390K pieces of content in Q1 2021 to 536.4K in Q2 2021 due to returning reviewer capacity. Restored content has continued to decrease from 367.5K pieces of content in Q2 2021 to 266.7K in Q3 2021 due to improved precision in our enforcement actions.</p> <p>Child Endangerment: Sexual Exploitation: Restored content increased significantly from 2.9K 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.</p> <p>Child Endangerment: Sexual Exploitation: 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.</p>
	Child Endangerment: Nudity and Physical Abuse	3.7k	800	19.2k	2.3k	700	167.2k	3k	800	21.1k	N/A	N/A	N/A	
	Child Endangerment: Sexual Exploitation	800	70	180.5k	700	30	2.8k	1k	30	2.8k	N/A	N/A	N/A	
Arms & Ammunition	Regulated Goods: Firearms	63k	10.7k	59.8k	44.1k	9k	60.7k	82.7k	20.6k	119.9k	125.3k	39.6k	128.1k	<p>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.</p>
	Violence and Incitement	361.8k	33.9k	198.9k	435.3k	37.9k	209.5k	N/A	N/A	N/A	N/A	N/A	N/A	



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

Authorized Metric: Appeals, Reinstatements

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

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary	
		Q4 2021			Q3 2021			Q2 2021			Q1 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		
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	311.4k	36.7k	262k	272.9k	32.9k	233.6k	536.4k	55.9k	311.9k	390k	73.1k	559.5k		
	Violence and Incitement	361.8k	33.9k	198.9k	435.3k	37.9k	209.5k	N/A	N/A	N/A	N/A	N/A	N/A	Adult Nudity and Sexual Activity: Appealed content has continued to increase from 390K pieces of content in Q1 2021 to 536.4K in Q2 2021 due to returning reviewer capacity. Restored content has continued to decrease from 367.5K pieces of content in Q2 2021 to 266.7K in Q3 2021 due to improved precision in our enforcement actions.	
	Violent and Graphic Content	3.8k	600	5.5k	3.3k	700	8k	3.7k	700	12.1k	3.7k	1k	10.5k	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.	
	Bullying and Harassment	799.4k	121.4k	242.5k	1m	149.8k	282.6k	919.2k	120.9k	210.8k	935.5k	112.6k	178.9k	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 Nudity and Sexual Exploitation	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	3.8k	300	46.7k	Child Endangerment: Sexual Exploitation: Restored content increased significantly from 2.9K 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	3.7k	800	19.2k	2.3k	700	167.2k	3k	800	21.1k	N/A	N/A	N/A	Child Endangerment: Sexual Exploitation: 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	800	70	180.5k	700	30	2.8k	1k	30	2.8k	N/A	N/A	N/A	Suicide and Self-Injury: 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. Restored content increased from 185.7K pieces of content in Q1 2021 to 279.8K in Q2 2021. In Q2, we took action on some content that wasn't violating, which we later restored. Restored content increased from 3.8K pieces of content in Q4 2020 to 185.7K in Q1 2021. This was due to an issue in February, when our media-matching technology removed a large amount of non-violating content, which we later restored.	
	Suicide and Self-Injury	200	50	95.2k	200	70	161.5k	300	60	279.7k	900	100	185.5k		
	Regulated Goods: Firearms	63k	10.7k	59.8k	44.1k	9k	60.7k	82.7k	20.6k	119.9k	125.3k	39.6k	128.1k		



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

Authorized Metric: Appeals, Reinstatements

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

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2021			Q3 2021			Q2 2021			Q1 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	
Death, Injury or Military Conflict	Violent and Graphic Content	3.8k	600	5.5k	3.3k	700	8k	3.7k	700	12.1k	3.7k	1k	10.5k	Suicide and Self-Injury: 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. Restored content increased from 185.7K pieces of content in Q1 2021 to 279.8K in Q2 2021. In Q2, we took action on some content that wasn't violating, which we later restored. Restored content increased from 3.8K pieces of content in Q4 2020 to 185.7K in Q1 2021. This was due to an issue in February, when our media-matching technology removed a large amount of non-violating content, which we later restored.
	Violence and Incitement	361.8k	33.9k	198.9k	435.3k	37.9k	209.5k	N/A	N/A	N/A	N/A	N/A	N/A	
	Suicide and Self Injury	200	50	95.2k	200	70	161.5k	300	60	279.7k	900	100	185.5k	
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 reinstated of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
Hate speech & acts of aggression	Hate Speech	768.8k	65.3k	227.7k	1.1m	90.7k	303k	1.4m	88.1k	328.1k	1.1m	84.8k	324.1k	Hate Speech: Restored content increased from 259.3K pieces of content in Q4 2020 to 408.7K in Q1 2021, primarily due to a technical issue in February that removed a large amount of non-violating content, which we then restored. 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	38.9k	5.2k	52.2k	64.7k	10.2k	85.2k	35.9k	3.5k	42.5k	38.8k	6.5k	63.2k	
	Dangerous Organizations: Organized Hate	35.7k	11.9k	115.4k	53.3k	23.8k	433.1k	87.8k	33.7k	484.7k	148.6k	45k	454.8k	
	Bullying and Harassment	799.4k	121.4k	242.5k	1m	149.8k	282.6k	919.2k	120.9k	210.8k	935.5k	112.6k	178.9k	



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

Authorized Metric: Appeals, Reinstatements

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

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2021			Q3 2021			Q2 2021			Q1 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	768.8k	65.3k	227.7k	1.1m	90.7k	303k	1.4m	88.1k	328.1k	1.1m	84.8k	324.1k	Hate Speech: Restored content increased from 259.3K pieces of content in Q4 2020 to 408.7K in Q1 2021, primarily due to a technical issue in February that removed a large amount of non-violating content, which we then restored.
	Bullying and Harassment	799.4k	121.4k	242.5k	1m	149.8k	282.6k	919.2k	120.9k	210.8k	935.5k	112.6k	178.9k	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	80k	27.7k	119.8k	43.9k	9.3k	83.5k	36.3k	6.1k	59.2k	36.4k	14.9k	97.1k	Regulated Goods: Drugs: Appealed content increased from 43.9K in Q3 2021 to 80K in Q4 2021, following a proportionate increase in content actioned. Restored content decreased from 112K pieces of content in Q1 2021 to 65.3K in Q2 2021 due to the adjustments we made to our proactive detection technology.
Spam or Harmful Content	Spam	21.6k	2k	54.1m	18.7k	1.2k	20.9m	14.5k	1.2k	30.6m	7.2k	1.2k	38.8m	Spam: 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	38.9k	5.2k	52.2k	64.7k	10.2k	85.2k	35.9k	3.5k	42.5k	38.8k	6.5k	63.2k	
	Dangerous Organizations: Organized Hate	35.7k	11.9k	115.4k	53.3k	23.8k	433.1k	87.8k	33.7k	484.7k	148.6k	45k	454.8k	
Debated Sensitive Social Issue	Hate Speech	768.8k	65.3k	227.7k	1.1m	90.7k	303k	1.4m	88.1k	328.1k	1.1m	84.8k	324.1k	Hate Speech: Restored content increased from 259.3K pieces of content in Q4 2020 to 408.7K in Q1 2021, primarily due to a technical issue in February that removed a large amount of non-violating content, which we then restored.
	Bullying and Harassment	799.4k	121.4k	242.5k	1m	149.8k	282.6k	919.2k	120.9k	210.8k	935.5k	112.6k	178.9k	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.



Question 1: How safe is the platform for consumers?

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2021	Q3 2021	Q2 2021	Q1 2021	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	0.02%-0.03%	0.02%-0.03%	0.03%	0.03%-0.04%	Adult Nudity and Sexual Activity: Prevalence remained relatively consistent across Q2 2021 and Q3 2021. 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: 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: This is the first time we are releasing prevalence for violence and incitement. During Q3 2021, it was about 0.02%. Prevalence remained relatively consistent across Q3 2021 and Q4 2021.
	Violence and Incitement	0.01%-0.02%	0.02%	N/A	N/A	



Question 1: How safe is the platform for consumers?

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2021	Q3 2021	Q2 2021	Q1 2021	
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	0.02%-0.03%	0.02%-0.03%	0.03%	0.03%-0.04%	<p>Adult Nudity and Sexual Activity: Prevalence remained relatively consistent across Q2 2021 and Q3 2021.</p> <p>Bullying and Harassment: Prevalence remained consistent across Q3 2021 and Q4 2021.</p> <p>Violence and Incitement: This is the first time we are releasing prevalence for violence and incitement. During Q3 2021, it was about 0.02%. Prevalence remained relatively consistent across Q3 2021 and Q4 2021.</p> <p>Violent and Graphic Content: Prevalence remained consistent across Q2 2021 and Q3 2021.</p>
	Violence and Incitement	0.01%-0.02%	0.02%	N/A	N/A	
	Violent and Graphic Content	0.01%-0.02%	0.01%-0.02%	0.01%	0.01%-0.02%	
	Bullying and Harassment	0.05%-0.06%	0.05%-0.06%	N/A	N/A	
	Child Nudity and Sexual Exploitation	N/A	N/A	N/A	Less than 0.05%	
	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 Instagram and measures the estimated percentage of those views that were of violating content.

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2021	Q3 2021	Q2 2021	Q1 2021	
Death, Injury or Military Conflict	Violent and Graphic Content	0.01%-0.02%	0.01%-0.02%	0.01%	0.01%-0.02%	Violent and Graphic Content: Prevalence remained consistent across Q2 2021 and Q3 2021. Violence and Incitement: This is the first time we are releasing prevalence for violence and incitement. During Q3 2021, it was about 0.02%. Prevalence remained relatively consistent across Q3 2021 and Q4 2021.
	Violence and Incitement	0.01%-0.02%	0.02%	N/A	N/A	Violence and Incitement: This is the first time we are releasing prevalence for violence and incitement. During Q3 2021, it was about 0.02%. Prevalence remained relatively consistent across Q3 2021 and Q4 2021.
	Suicide and Self Injury	Less than 0.05%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	We do not report prevalence of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	
Hate speech & acts of aggression	Hate Speech	0.02%-0.03%	0.02%	N/A	N/A	Hate Speech: This is the first time we are releasing prevalence for hate speech. During Q3 2021, it was about 0.02%. Prevalence remained relatively consistent across Q3 2021 and Q4 2021. Dangerous Organizations: Organized Hate-We cannot estimate prevalence for organized hate right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data. Bullying and Harassment: Prevalence remained consistent across Q3 2021 and Q4 2021.
	Dangerous Organizations: Terrorism	Less than 0.06%	Less than 0.05%	Less than 0.05%	Less than 0.05%	
	Dangerous Organizations: Organized Hate	N/A	N/A	N/A	N/A	
	Bullying and Harassment	0.05%-0.06%	0.05%-0.06%	N/A	N/A	



Question 1: How safe is the platform for consumers?

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

GARM Category	Relevant Policy	Latest Period		Previous Period		Commentary
		Q4 2021	Q3 2021	Q2 2021	Q1 2021	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	0.02%-0.03%	0.02%	N/A	N/A	Hate Speech: This is the first time we are releasing prevalence for hate speech. During Q3 2021, it was about 0.02%. Bullying and Harassment: Prevalence remained consistent across Q3 2021 and Q4 2021.
	Bullying and Harassment	0.05%-0.06%	0.05%-0.06%	N/A	N/A	
Illegal Drugs/Tobacco/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 0.06%	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%-0.03%	0.02%	N/A	N/A	Hate Speech: This is the first time we are releasing prevalence for hate speech. During Q3 2021, it was about 0.02%. Prevalence remained relatively consistent across Q3 2021 and Q4 2021. Bullying and Harassment: Prevalence remained consistent across Q3 2021 and Q4 2021.
	Bullying and Harassment	0.05%-0.06%	0.05%-0.06%	N/A	N/A	



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

Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Instagram

GARM Category	Relevant Policy	Latest Period				Previous Period			
		Q4 2021		Q3 2021		Q2 2021		Q1 2021	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	11.3m	94.3%	10.9m	95.4%	9.4m	96.7%	10.4m	96.2%
	Child Endangerment: Sexual Exploitation	2.6m	97.3%	1.6m	96.3%	1.4m	96.3%	N/A	N/A
	Child Endangerment: Nudity and Physical Abuse	983.4k	95.3%	526.5k	92.3%	458.1k	95.8%	N/A	N/A
Arms & Ammunition	Regulated Goods: Firearms	195k	94.3%	154.4k	95.8%	76.3k	90.6%	90.4k	89.7%
	Violence and Incitement	2.6m	96%	3.3m	96.4%	N/A	N/A	N/A	N/A
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	11.3m	94.3%	10.9m	95.4%	9.4m	96.7%	10.4m	96.2%
	Violence and Incitement	2.6m	96%	3.3m	96.4%	N/A	N/A	N/A	N/A
	Violent and Graphic Content	5.5m	98.7%	10.7m	99.3%	7.6m	98.8%	5.5m	98.4%
	Bullying and Harassment	6.6m	82.1%	7.8m	83.2%	4.5m	71.5%	5.5m	78.6%
	Child Nudity and Sexual Exploitation	N/A	N/A	N/A	N/A	N/A	N/A	812.3k	98.1%
	Child Endangerment: Nudity and Physical Abuse	983.4k	95.3%	526.5k	92.3%	1.4m	96.3%	N/A	N/A
	Child Endangerment: Sexual Exploitation	2.6m	97.3%	1.6m	96.3%	458.1k	95.8%	N/A	N/A
	Suicide and Self-Injury	7.8m	98.4%	3.5m	96.8%	3m	96%	2.7m	94.5%
	Regulated Goods: Firearms	195k	94.3%	154.4k	95.8%	76.3k	90.6%	90.4k	89.7%
Death, Injury or Military Conflict	Violent and Graphic Content	5.5m	98.7%	10.7m	99.3%	7.6m	98.8%	5.5m	98.4%
	Violence and Incitement	2.6m	96%	3.3m	96.4%	N/A	N/A	N/A	N/A
	Suicide and Self Injury	7.8m	98.4%	3.5m	96.8%	3m	96%	2.7m	94.5%
Online piracy	Intellectual Property: Copyright	866.1k	88%	778.3k	86%	888.3k	89.1%	859.6k	91.5%
	Intellectual Property: Counterfeit	302.4k	93%	317.5k	94%	324.8k	94.3%	253.9k	93.5%
	Intellectual Property: Trademark	154.6k	71%	144.8k	69%	127.9k	55.6%	201.6k	69.8%



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

Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Instagram

GARM Category	Relevant Policy	Latest Period				Previous Period			
		Q4 2021		Q3 2021		Q2 2021		Q1 2021	
		Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive	Content Removed	% Proactive
Hate speech & acts of aggression	Hate Speech	3.8m	91.9%	6m	93.8%	9.8m	95.1%	6.3m	93.4%
	Dangerous Organizations: Terrorism	905.3k	79.5%	685.2k	72.3%	336.9k	99.7%	429k	99.6%
	Dangerous Organizations: Organized Hate	332.2k	84.8%	305.8k	82.7%	367.3k	77.7%	324.6k	77.3%
	Bullying and Harassment	6.6m	82.1%	7.8m	83.2%	4.5m	71.5%	5.5m	78.6%
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	3.8m	91.9%	6m	93.8%	9.8m	95.1%	6.3m	93.4%
	Bullying and Harassment	6.6m	82.1%	7.8m	83.2%	4.5m	71.5%	5.5m	78.6%
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	1.2m	95%	1.8m	97.4%	1.1m	95.4%	1m	91.4%
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	905.3k	79.5%	685.2k	72.3%	336.9k	99.7%	429k	99.6%
	Dangerous Organizations: Organized Hate	332.2k	84.8%	305.8k	82.7%	367.3k	77.7%	324.6k	77.3%
Debated Sensitive Social Issue	Hate Speech	3.8m	91.9%	6m	93.8%	9.8m	95.1%	6.3m	93.4%
	Bullying and Harassment	6.6m	82.1%	7.8m	83.2%	4.5m	71.5%	5.5m	78.6%



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
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	
	Child Endangerment: Sexual Exploitation	Content actioned for child sexual exploitation has remained relatively consistent from 1.4 million in Q2 2021 and 1.6 million Q3 2021. 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 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 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 95.8% 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	This is the first time we are releasing content actioned for violence and incitement. During Q3 2021, we actioned 3.3 million pieces of content.
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	
	Violence and Incitement	This is the first time we are releasing content actioned for violence and incitement. During Q3 2021, we actioned 3.3 million pieces of content. Content actioned decreased from 3.3 million in Q3 2021 to 2.6 million in Q4 2021 due to a decrease in overall and violating comments.
	Violent and Graphic 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: Content actioned increased from 4.5 million to 7.8 million as we expanded our proactive detection technology to more languages.
	Child Nudity and Sexual Exploitation	Child Nudity and Sexual Exploitation: From Q2 2021 we've added more data and 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	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	Content actioned for child sexual exploitation has remained relatively consistent from 1.4 million in Q2 2021 and 1.6 million Q3 2021. 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 for suicide and self-injury increased from 3 million in Q2 2021 to 3.5 million in Q3 2021, in part due to recent improvements in our text and image detection technology.
Death, Injury or Military Conflict	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 95.8% in Q3 2021 due to improved and expanded proactive detection technologies.
	Violent and Graphic 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	This is the first time we are releasing content actioned for violence and incitement. During Q3 2021, we actioned 3.3 million pieces of content. Content actioned decreased from 3.3 million in Q3 2021 to 2.6 million in Q4 2021 due to a decrease in overall and violating comments.
	Suicide and Self-Injury	Content actioned for suicide and self-injury increased from 3 million in Q2 2021 to 3.5 million in Q3 2021, in part due to recent improvements in our text and image detection technology. Content actioned increased from 3.5 million to 7.8 million due to the expansion of our media-matching technology to identify and remove old, violating content.



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. 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.
	Intellectual Property: Counterfeit	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: 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 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 for organized hate decreased from 367.4K in Q3 2021 to 305.7K in Q3 2021 returning back to pre-Q2 levels after being elevated in Q2 2021 due to offline events. 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: 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: Content actioned increased from 4.5 million to 7.8 million as we expanded our proactive detection technology to more languages.v
Illegal Drugs/Tobacco/cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	Content actioned increased from 1.1 million in Q2 2021 to 1.8 million in Q3 2021 due to improved proactive detection technologies we launched in late Q2 2021. 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 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 for organized hate decreased from 367.4K in Q3 2021 to 305.7K in Q3 2021 returning back to pre-Q2 levels after being elevated in Q2 2021 due to offline events. 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: 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

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2021			Q3 2021			Q2 2021			Q1 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	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	0	0	227.8k	0	0	159.2k	0	0	149.7k	0	10	264.4k	<p>The number of appeals is 0 because we couldn't offer them due to the impact of COVID-19 on our reviewers. We let users know about this, and if they felt we made a mistake, we still gave people the option to tell us they disagreed with our decision. We reviewed many of these instances, and restored content when appropriate.</p> <p>Adult Nudity and Sexual Activity: 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.</p> <p>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.</p>
	Child Endangerment: Sexual Exploitation	0	0	1.6k	0	0	300	0	0	300	N/A	N/A	N/A	
	Child Endangerment: Nudity and Physical Abuse	0	0	13.6k	0	0	168.3k	0	0	4.5k	N/A	N/A	N/A	
Arms & Ammunition	Regulated Goods: Firearms	0	0	21.7k	0	0	7.8k	0	0	1.8k	0	0	1.4k	<p>The number of appeals is 0 because we couldn't offer them due to the impact of COVID-19 on our reviewers. We let users know about this, and if they felt we made a mistake, we still gave people the option to tell us they disagreed with our decision. We reviewed many of these instances, and restored content when appropriate.</p> <p>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.</p>
	Violence and Incitement	0	0	49.9k	0	0	21.4k	N/A	N/A	N/A	N/A	N/A	N/A	



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

Authorized Metric: Appeals, Reinstatements

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

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2021			Q3 2021			Q2 2021			Q1 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	
Crime & Harmful acts to individuals and Society, Human Right Violations	Adult Nudity and Sexual Activity	0	0	227.8k	0	0	159.2k	0	0	149.7k	0	10	264.4k	The number of appeals is 0 because we couldn't offer them due to the impact of COVID-19 on our reviewers. We let users know about this, and if they felt we made a mistake, we still gave people the option to tell us they disagreed with our decision. We reviewed many of these instances, and restored content when appropriate.
	Violence and Incitement	0	0	49.9k	0	0	21.4k	N/A	N/A	N/A	N/A	N/A	N/A	
	Violent and Graphic Content	0	0	2.5m	0	0	21.2k	0	0	10.6k	0	0	10.7k	Adult Nudity and Sexual Activity: 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.
	Bullying and Harassment	0	0	182.1k	0	0	91.5k	0	0	21.4k	0	0	52.7k	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.
	Child Nudity and Sexual Exploitation	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0	0	3.5k	Violent and Graphic Content: Restored content increased from 10.6K pieces of content in Q2 2021 to 21.2K in Q3 2021 primarily as we restored some viral disturbing images but which were shared in the context of raising awareness. 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.
	Child Endangerment: Nudity and Physical Abuse	0	0	13.6k	0	0	168.3k	0	0	4.5k	N/A	N/A	N/A	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 Endangerment: Sexual Exploitation	0	0	1.6k	0	0	300	0	0	300	N/A	N/A	N/A	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.
	Suicide and Self-Injury	0	0	600	0	0	10.6k	0	0	79.2k	0	0	119.7k	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.
	Regulated Goods: Firearms	0	0	21.7k	0	0	7.8k	0	0	1.8k	0	0	1.4k	Suicide and Self-Injury: 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.



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

Authorized Metric: Appeals, Reinstatements

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

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2021			Q3 2021			Q2 2021			Q1 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	
Death, Injury or Military Conflict	Violent and Graphic Content	0	0	2.5m	0	0	21.2k	0	0	10.6k	0	0	10.7k	The number of appeals is 0 because we couldn't offer them due to the impact of COVID-19 on our reviewers. We let users know about this, and if they felt we made a mistake, we still gave people the option to tell us they disagreed with our decision. We reviewed many of these instances, and restored content when appropriate.
	Violence and Incitement	0	0	49.9k	0	0	21.4k	N/A	N/A	N/A	N/A	N/A	N/A	Violent and Graphic Content: Restored content increased from 10.6K pieces of content in Q2 2021 to 21.2K in Q3 2021 primarily as we restored some viral disturbing images but which were shared in the context of raising awareness. 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.
	Suicide and Self Injury	0	0	600	0	0	10.6k	0	0	79.2k	0	0	119.7k	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.
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	We do not report content appealed and reinstated of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
Hate speech & acts of aggression	Hate Speech	0	0	63.6k	0	0	43.1k	0	0	30.8k	0	0	43.8k	The number of appeals is 0 because we couldn't offer them due to the impact of COVID-19 on our reviewers. We let users know about this, and if they felt we made a mistake, we still gave people the option to tell us they disagreed with our decision. We reviewed many of these instances, and restored content when appropriate.
	Dangerous Organizations: Terrorism	0	0	4k	0	0	3.5k	0	0	1.1k	0	0	60	
	Dangerous Organizations: Organized Hate	0	0	7.5k	0	0	3.7k	0	0	2.2k	0	0	6.5k	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	0	0	182.1k	0	0	91.5k	0	0	21.4k	0	0	52.7k	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.



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

Authorized Metric: Appeals, Reinstatements

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

GARM Category	Relevant Policy	Latest Period						Previous Period						Commentary
		Q4 2021			Q3 2021			Q2 2021			Q1 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	0	0	63.6k	0	0	43.1k	0	0	30.8k	0	0	43.8k	The number of appeals is 0 because we couldn't offer them due to the impact of COVID-19 on our reviewers. We let users know about this, and if they felt we made a mistake, we still gave people the option to tell us they disagreed with our decision. We reviewed many of these instances, and restored content when appropriate.
	Bullying and Harassment	0	0	182.1k	0	0	91.5k	0	0	21.4k	0	0	52.7k	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: 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	0	0	27.6k	0	0	35.6k	0	0	19.6k	0	0	28.5k	The number of appeals is 0 because we couldn't offer them due to the impact of COVID-19 on our reviewers. We let users know about this, and if they felt we made a mistake, we still gave people the option to tell us they disagreed with our decision. We reviewed many of these instances, and restored content when appropriate.
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	0	0	4k	0	0	3.5k	0	0	1.1k	0	0	60	The number of appeals is 0 because we couldn't offer them due to the impact of COVID-19 on our reviewers. We let users know about this, and if they felt we made a mistake, we still gave people the option to tell us they disagreed with our decision. We reviewed many of these instances, and restored content when appropriate.
	Dangerous Organizations: Organized Hate	0	0	7.5k	0	0	3.7k	0	0	2.2k	0	0	6.5k	
Debated Sensitive Social Issue	Hate Speech	0	0	63.6k	0	0	43.1k	0	0	30.8k	0	0	43.8k	The number of appeals is 0 because we couldn't offer them due to the impact of COVID-19 on our reviewers. We let users know about this, and if they felt we made a mistake, we still gave people the option to tell us they disagreed with our decision. We reviewed many of these instances, and restored content when appropriate.
	Bullying and Harassment	0	0	182.1k	0	0	91.5k	0	0	21.4k	0	0	52.7k	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: 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 drives the work we do at Twitter and underpins our efforts to serve the public conversation. Over the past year, we've continued to protect this conversation as we experienced and continue to navigate severe global challenges, including the coronavirus pandemic. We've also seen concerted attempts by governments to limit access to the Internet generally and to Twitter specifically.

The metrics in this GARM report reflect the enforcement of the Twitter Rules, which apply to everyone who uses Twitter. Our rules exist to ensure people can participate in the public conversation freely and safely.

Our Brand Safety Policies, as well as the controls we offer people and advertisers, build upon the foundation laid by the Twitter Rules to promote a safe advertising experience for all users and brands, and inform the context in which we serve ads. We look forward to providing additional transparency for advertisers through partnerships with third parties like DoubleVerify (DV) and Integrated Ad Science (IAS) who are building independent brand safety reporting solutions that will provide additional insights aligned with the GARM framework. For purposes of this report, however, the metrics we share pertain specifically to enforcement of the Twitter Rules.

Our latest Twitter Transparency Center update includes data from January 1 2021 to June 30th 2021. While we continue to share data across consistent, recurring categories, we're always looking to incorporate new data that can provide meaningful insights into the impact of our actions.

First introduced in our Transparency Report for the second half of 2020, our impressions metric provides meaningful transparency on the number of views a violative Tweet received prior to removal. In total, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets globally, from January 1 through June 30, 2021. During this time period, Twitter removed 4.7 million Tweets that violated the Twitter Rules; 68% of which 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.

More broadly, as we work to remove harmful, violative content quickly and at scale, whether amid a global health crisis or during a humanitarian crisis, these numbers represent both our present efficiency and where improvement is needed. Our goal is to improve these numbers over time, taking enforcement action on violative content before it's even viewed.

As the COVID-19 pandemic evolves around the world, we continue to take enforcement action on misleading information about COVID-19, and amplify the most current, up-to-date, and authoritative information on the situation as it unfolds.

Additionally, Civic Integrity policy enforcements decreased significantly in H1 2021, by 91%, compared to the previous reporting period, due to the end of the US 2020 election cycle. During the United States 2020 election, we enacted a set of policy, enforcement and product changes to add context, encourage thoughtful consideration, and reduce the potential for misleading information to spread on Twitter.

In addition to labels, we are continuing to increase our use of machine learning and automation to take a wide range of actions on potentially misleading and manipulative content, and proactively surface abusive content for review. Like many organizations – both public and private around the world – the disruptions caused by COVID-19 made an impact on our company and are reflected in some of the data shared in this report. We're committed to enabling safe and healthy conversations on the service, and we're always looking for ways to share more context about our enforcement of the Twitter Rules.

We also aim to build safety into Twitter as a platform through third-party partnerships and new product features. During the first half of 2021 (the period covered by the most recent data included in this report), we successfully earned the Trustworthy Accountability Group (TAG) Brand Safety Certified Seal, and we engaged OpenSlate, a DoubleVerify company, to provide third-party verification of the safety and suitability of the content in our Twitter Amplify offering. OpenSlate's analysis found that of the more than 455,000 monetized videos analyzed, 100% fell above the GARM Brand Safety Floor. We also released important health & safety product features in the first half of 2021, including conversation settings for advertisers, and prompts that encourage people to pause and reconsider a potentially harmful or offensive reply before they hit send.

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. Additionally, we're looking forward to 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 we're proud of the progress made in partnership with GARM on this front towards greater industry-wide visibility and transparency.

April 2022



Question 1: How safe is the platform for consumers?

Next Best Measure: Impressions

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

GARM Category	Relevant Twitter Policy	Latest Period – H1 2021	Previous Period – H2 2020	Commentary
		Impressions	Impressions	
Adult & explicit sexual content	Non-consensual nudity	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.	In total, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets globally, from July 1, 2020 through December 31 2020.	<p>During January 2021 - June 2021, Twitter removed 4.7 million Tweets (a 24% increase over the previous period) that violated the Twitter Rules.</p> <ul style="list-style-type: none"> • 68% of these Tweets received fewer than 100 impressions prior to removal. • 24% received between 100 and 1,000 impressions prior to removal. • Only 8% of removed Tweets had more than 1,000 impressions prior to removal. <p>We do not report breakdowns by policy for this metric.</p>
	Sensitive media			
	Child sexual exploitation			
Arms & ammunition	Illegal or certain regulated goods or services			
Crime & harmful acts to individuals and society, human right violations	Violence			
	Abuse/harassment			
Death, injury or military conflict	Promoting suicide or self-harm			
Online piracy	Copyright			
	Trademark			
Hate speech & acts of aggression	Hateful conduct			
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media			
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services			
Spam or harmful content	Private information			
	Impersonation			
	Platform manipulation			
Terrorism	Terrorism/violent extremism			
Debated sensitive social issues	N/A			
Other	Civic integrity			
	COVID-19 misleading information			



Question 2: How safe is the platform for advertisers?

Next best measure: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2021			Previous Period – H2 2020			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Adult & explicit sexual content	Non-consensual nudity	29,635	7,519	64,596	27,087	3,693	52,442	<p>In December of 2020, Twitter announced partnerships with DoubleVerify (DV) and Integral Ad Science (IAS) to provide independent reporting on the context in which ads appear on Twitter. Since then, we have made significant progress in developing third-party brand safety measurement solutions with DV and IAS. We spent 2021 working closely with each partner to build and refine the technical solutions necessary to enable them to measure and classify the content that appears directly adjacent to a brand's Promoted Tweets.</p> <p>In early 2022, we moved into the internal testing phase of these integrations and expect to begin Beta testing with a small group of hand-selected advertisers in the coming months.</p> <p>Additionally, we remain dedicated to building partnerships that further our mission to protect and serve the public conversation:</p> <ul style="list-style-type: none"> - Last year, we shared that we reached an agreement with the Media Ratings Council (MRC) to initiate our pre-assessment for their Brand Safety accreditation. Since that announcement, we worked closely with the MRC to answer questions and provide detailed information about our operations, methodology, processing, reporting and disclosures in the brand safety space. Twitter is looking forward to reviewing the pre-assessment summary with MRC and continuing to work towards the MRC Brand Safety accreditation. - We're also excited to share that Twitter has earned the Trustworthy Accountability Group (TAG) Brand Safety Certified Seal for the second time since its inception in 2021. This seal covers Twitter's global operations. For this year's recertification, we have newly incorporated TAG's anti-piracy standards to mitigate the spread of pirated content online.
	Sensitive media	1,630,554	164,260	1,655,608	706,979	42,801	728,778	
	Child sexual exploitation	456,146	453,754	6,087	469,439	464,804	9,178	
Arms & ammunition	Illegal or certain regulated goods or services	175,798	87,530	420,950	103,285	53,696	236,119	
Crime & harmful acts to individuals and society, human right violations	Violence	89,245	66,445	101,907	49,146	34,829	59,933	
	Abuse/harassment	1,043,525	99,565	1,547,654	964,459	86,202	1,448,418	
Death, injury or military conflict	Promoting suicide or self-harm	345,100	8,621	413,769	188,561	4,287	226,905	
Online piracy	Copyright	Notices Issued: 171,747	Accounts Affected: 796,506	Tweets Withheld: 432,759	Notices Issued: 129,880	Accounts Affected: 675,999	Tweets Withheld: 845,183	
	Trademark	Trademark Notices: 20,121			Trademark Notices: 16,193			
Hate speech & acts of aggression	Hateful conduct	1,108,722	133,585	1,606,979	1,126,990	157,815	1,628,281	
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	1,630,554	164,260	1,655,608	706,979	42,801	728,778	
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services	175,798	87,530	420,950	103,285	53,696	236,119	
Spam or harmful content	Private information	30,714	3,178	54,590	42,894	2,885	65,601	
	Impersonation	216,846	199,229	21,188	141,033	126,750	15,816	
	Platform manipulation	Anti-Spam Challenges Issued: 130,289,899			Anti-Spam Challenges Issued: 143,211,618			
Terrorism	Terrorism/violent extremism	44,974	44,974		58,750	58,750		
Debated sensitive social issues	N/A							
Other	Civic integrity	581	23	593	6,469	64	8,122	
	COVID-19 misleading information	27,935	617	33,761	3,399	597	3,846	



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

Authorized Metric: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2021			Previous Period – H2 2020			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Adult & explicit sexual content	Non-consensual nudity	29,635	7,519	64,596	27,087	3,693	52,442	<p>There was a 9% increase in the number of accounts actioned for violations of our non-consensual nudity policy during this reporting period.</p> <p>This reporting period saw the largest increase in the number of accounts suspended under this policy. We launched initiatives to better detect and take action on content, which led to an increase in accounts suspended under our non-consensual nudity policy by 104%.</p>
	Sensitive media	1,630,554	164,260	1,655,608	706,979	42,801	728,778	<p>There was a 131% increase in the number of accounts actioned for violations of our sensitive media policy during this reporting period.</p> <p>We saw the largest increase in the number of accounts actioned and content removed during this reporting period. Initiatives were launched to bolster operational capacity that resulted in an increase in actioning of content that violates our sensitive media policies.</p>
	Child sexual exploitation	456,146	453,754	6,087	469,439	464,804	9,178	<p>There was a 3% decrease in the number of accounts actioned for violations of our child sexual exploitation policy during this reporting period.</p>
Arms & ammunition	Illegal or certain regulated goods or services	175,798	87,530	420,950	103,285	53,696	236,119	<p>There was an 70% increase in the number of accounts actioned for violations of our illegal or certain regulated goods or services policy during this reporting period.</p> <p>Since the launch of the policy in 2019, and more specifically at the end of the last year, we have continued to refine our enforcement guidelines. This improvement resulted in more accounts being actioned for violation of the policy which in turn triggered an increase in the number of accounts trying to circumvent their previous suspension or enforcement action, thus violating Twitter policy on ban evasion.</p>
Crime & harmful acts to individuals and society, human right violations	Violence	89,245	66,445	101,907	49,146	34,829	59,933	<p>There was an 82% increase in the number of accounts actioned for violations of our violence policies this reporting period.</p> <p>Our policies prohibit sharing of content that threatens violence against an individual or a group of people. We also prohibit the glorification of violence. We saw a significant increase in the number of content removed for violence and accounts suspended due to initiatives launched to bolster operational capacity.</p>
	Abuse/harassment	1,043,525	99,565	1,547,654	964,459	86,202	1,448,418	<p>There was an 8% increase in the number of accounts actioned for violations of our abuse policy during this reporting period.</p> <p>In this reporting period, we updated our policy to remove the targeting requirement for content that denies that mass murder or other mass casualty events took place, where we can verify that the event occurred, and when the content is shared with abusive intent.</p>



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

Authorized Metric: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2021			Previous Period – H2 2020			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Death, injury or military conflict	Promoting suicide or self-harm	345,100	8,621	413,769	188,561	4,287	226,905	<p>There was an 83% increase in the number of accounts actioned for violations of our suicide or self-harm policy during this reporting period.</p> <p>During this reporting period there was a significant increase in the volume of accounts actioned (83%), accounts suspended (101%), and content removed (82%). Initiatives were launched to better detect and take action on content that violated our policy on suicide and self-harm which led to the spike in enforcement numbers.</p>
Online piracy	Copyright	Notices Issued: 171,747	Accounts Affected: 796,506	Tweets Withheld: 432,759	Notices Issued: 129,880	Accounts Affected: 675,999	Tweets Withheld: 845,183	<p>We report on DMCA takedown notices submitted through our web form or otherwise sent to Twitter, such as via fax or mail.</p> <p>We saw a 6% increase in DMCA takedown notices submitted, and a 17% increase in accounts affected. Tweets withheld dropped by 49% while media withheld increased by 18%, as Twitter's operations were affected due to the unprecedented COVID-19 pandemic.</p> <p>We provide affected account holders with a copy of the related DMCA takedown notice when their media or Tweets are withheld. The notification includes instructions on how to file a counter-notice (in case they believed the content was removed in error) and also how to seek a retraction from the original reporter.</p>
	Trademark	Trademark Notices: 20,121			Trademark Notices: 16,193			Twitter received 24% more trademark notices, affecting 33% more accounts since our last report.
Hate speech & acts of aggression	Hateful conduct	1,108,722	133,585	1,606,979	1,126,990	157,815	1,628,281	<p>There was a 2% decrease in the number of accounts actioned for violations of our hateful conduct policy during this reporting period.</p> <p>Our Hateful Conduct policy was updated in January 2021 to expand our enforcement approach towards content that incites others to discriminate by denying support to the economic enterprise of an individual or group because of their perceived membership in a protected category. In addition to the policy update, we also removed the targeting requirement for content aimed at individuals or groups that references forms of violence or violent events where a protected category was the primary target or victims and where the intent is to harass.</p>
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	1,630,554	164,260	1,655,608	706,979	42,801	728,778	<p>There was a 131% increase in the number of accounts actioned for violations of our sensitive media policy during this reporting period.</p> <p>We saw the largest increase in the number of accounts actioned and content removed during this reporting period. Initiatives were launched to bolster operational capacity that resulted in an increase in actioning of content that violates our sensitive media policies.</p>



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

Authorized Metric: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2021			Previous Period – H2 2020			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Spam or harmful content	Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	175,798	87,530	420,950	103,285	53,696	236,119	<p>There was an 70% increase in the number of accounts actioned for violations of our illegal or certain regulated goods or services policy during this reporting period.</p> <p>Since the launch of the policy in 2019, and more specifically at the end of the last year, we have continued to refine our enforcement guidelines. This improvement resulted in more accounts being actioned for violation of the policy which in turn triggered an increase in the number of accounts trying to circumvent their previous suspension or enforcement action, thus violating Twitter policy on ban evasion.</p>
	Private information	30,714	3,178	54,590	42,894	2,885	65,601	<p>There was a 28% decrease in the number of accounts actioned for violations of our private information policy during this reporting period.</p>
	Impersonation	216,846	199,229	21,188	141,033	126,750	15,816	<p>There was a 54% increase in the number of accounts actioned for violations of our impersonation policy during this reporting period.</p> <p>This reporting period saw more activity related to impersonation scams from accounts based in West Africa and Southeast Asia, which may account for the increase in accounts actioned under our impersonation policy.</p>
	Platform manipulation	Anti-Spam Challenges Issued: 130,289,899			Anti-Spam Challenges Issued: 143,211,618			<p>One way we fight manipulation and spam at scale is to use anti-spam challenges to confirm whether an authentic account holder is in control of accounts engaged in suspicious activity. For example, we may require the account holder to verify a phone number or email address, or to complete a reCAPTCHA test. These challenges are simple for authentic account owners to solve, but difficult (or costly) for spammers to complete. Accounts which fail to complete a challenge within a specified period of time may be suspended.</p> <p>These anti-spam challenges decreased by approximately 9% compared to the previous reporting period. We believe this can be attributed to ongoing efforts to reduce the impact of anti-spam challenges on legitimate users during this reporting period.</p>



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

Authorized Metric: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Twitter Policy	Latest Period – H1 2021			Previous Period – H2 2020			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Terrorism	Terrorism/violent extremism	44,974	44,974		58,750	58,750		<p>There was a 23% decrease in the number of accounts actioned for violations of our terrorism / violent extremism policy during this reporting period. Of those accounts, 93% were proactively identified and actioned.</p> <p>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).</p>
Debated sensitive social issues	N/A							
Other	Civic integrity	581	23	593	6,469	64	8,122	<p>There was a 91% decrease in the number of accounts actioned for violations of our civic integrity policy during this reporting period.</p> <p>The end of the 2020 US election cycle led to a significant decrease in the number of accounts actioned under our civic integrity policy since the last report.</p>
	COVID-19 misleading information	27,935	617	33,761	3,399	597	3,846	<p>There was a 722% increase in the number of accounts actioned for violations of our COVID-19 misleading information policy during this reporting period. This number does not include accounts where we applied a label or warning message.</p> <p>Since the introduction of COVID-19 guidance last year, there was increased focus on scaling the enforcement of the policy in particular in areas related to vaccine misinformation. In instances where accounts repeatedly violate this policy, a strike system is now used to determine if further enforcement actions should be applied. These actions include requests for tweet deletion, temporary account locks and permanent suspensions. We believe this system further helps to reduce the spread of potentially harmful and misleading information on Twitter, particularly for high-severity violations of our rules.</p>



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

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	<p>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.</p> <p>In H2 2020, we used technology to proactively surface 65% of the abusive content actioned for human review, instead of relying on reports from people using Twitter. This is a metric that we report at our discretion, and we may not disclose this metric every reporting period. H2 2020 is the latest period for which we've published this metric.</p>
	Sensitive media	
	Child sexual exploitation	
Arms & ammunition	Illegal or certain regulated goods or services	
Crime & harmful acts to individuals and society, human right violations	Violence	
	Abuse/harassment	
Death, injury or military conflict	Promoting suicide or self-harm	
Online piracy	Copyright	
	Trademark	
Hate speech & acts of aggression	Hateful conduct	
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services	
Spam or harmful content	Private information	
	Impersonation	
	Platform manipulation	
Terrorism	Terrorism/violent extremism	
Debated sensitive social issues	N/A	
Other	Civic integrity	
	COVID-19 misleading information	



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

Not submitted

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

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

Overview

We have continued to expand the information we provide in each report, and since the start of 2021 we have added insight into the volume of content removed at zero views, accounts removed from the full TikTok experience on the suspicion of being under the age of 13, and fake engagement. Starting with our Q4 report, we are providing information about content removals in more markets and ongoing improvements to our systems which aim to detect, flag, and, in some cases, remove violative content. These investments have helped meaningfully improve the speed at which we identify and remove violations of our harassment and bullying and hateful behavior policies in particular.

Automated Detection Removals

We continue to expand our system that detects and removes certain categories of violations at upload – including adult nudity and sexual activities, minor safety, and illegal activities and regulated goods. As a result, the volume of automated removals has increased, which improves the overall safety of our platform and enables our team to focus more time on reviewing contextual or nuanced content, such as hate speech, bullying and harassment, and misinformation.

From our first enforcement report in 2021 to our most recent report, we have steadfastly made progress on removing violations before they receive a single view. For instance, from Q1 to Q4 2021, removals of content at zero views improved by 14.7% for harassment and bullying content, 10.9% for hateful behavior, 16.2% for violent extremism, and 7.7% for suicide, self-harm, and dangerous acts.

Expanding our policy to protect the security, integrity, availability, and reliability of our platform

This includes prohibiting unauthorized access to TikTok, as well as TikTok content, accounts, systems, or data, and prohibiting the use of TikTok to perpetrate criminal activity. In addition to educating our community on ways to spot, avoid, and report suspicious activity, we are opening state-of-the-art cyber incident monitoring and investigative response centers in Washington DC, Dublin, and Singapore this year. TikTok’s Fusion Center operations enable follow-the-sun threat monitoring and intelligence gathering, as we continue working with industry-leading experts to test and enhance our defenses.

TikTok

Adding clarity on the types of hateful ideologies prohibited on our platform

This includes deadnaming, misgendering, or misogyny as well as content that supports or promotes conversion therapy programs. Though these ideologies have long been prohibited on TikTok, we have heard from creators and civil society organizations that it's important to be explicit in our Community Guidelines. On top of this, we hope our recent feature enabling people to add their pronouns will encourage respectful and inclusive dialogue on our platform

Strengthening our dangerous acts and challenges policy.

We continue to enact the stricter approach we previously announced to help prevent such content - including suicide hoaxes - from spreading on our platform. This previously sat within our suicide and self-harm policies but will now be highlighted in a separate policy category with more detail so it's even easier for our community to familiarize themselves with these guidelines. As part of our ongoing work to help our community understand online challenges and stay safe while having fun, we have worked with experts to launch new videos from creators that call on our community to follow four helpful steps when assessing content online - stop, think, decided and act. Community members can also view these videos at our #SaferTogether hub on the Discover page.

Fostering authentic engagement in comments

Alongside our work to proactively remove abusive and hateful content or behavior that violates our Community Guidelines, we also continue to explore new ways to help our community feel more in control over comments. We have started testing a way to let individuals identify comments they believe to be irrelevant or inappropriate. This community feedback will add to the range of factors we already use to help keep the comment section consistently relevant and a place for genuine engagement. To avoid creating ill-feeling between community members or demoralize creators, only the person who registered a dislike on a comment will be able to see that they have done so.

Finally

There's no finish line when it comes to keeping people safe, and our latest report and continued safety improvements reflect our ongoing 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



Question 1: How safe is the platform for consumers?

Next best measure: Videos removed by Policy Violation

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

Q3: **91,445,802** videos removed, Q4: **85,794,222** videos removed

GARM Category	Relevant Policy	Q4 2021	Q3 2021	Previous Period – Q2 2021	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	45.1%	51%	41.2%	Minor Safety includes anything that may perpetuate the abuse, harm, endangerment, or exploitation of minors
	Adult nudity and sexual activities	10.9%	11%	13.99%	
Arms & Ammunition	Illegal activities and regulated goods - Weapons	19.5%	16.6%	20.8%	Figure represents all Illegal activities and regulated goods, which includes <ul style="list-style-type: none"> Arms & Ammunition Crime & Harmful acts to individuals and Society, Human Right Violations Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol
Crime & Harmful acts to individuals and Society, Human Right Violations	Illegal activities and regulated goods - Criminal Activities	19.5%	16.6%	20.8%	Figure represents all Illegal activities and regulated goods, which includes <ul style="list-style-type: none"> Arms & Ammunition Crime & Harmful acts to individuals and Society, Human Right Violations Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol
Death, Injury or Military Conflict	Violent and graphic content	8.5%	7.4%	7.7%	
Online piracy	Integrity and authenticity - Intellectual property violations	0.6%	0.5%	0.7%	Figure Represents all Integrity and Authenticity, including: <ul style="list-style-type: none"> Online Piracy Spam This figure also includes videos removed for Misinformation, as defined by TikTok Community Guidelines



Question 1: How safe is the platform for consumers?

Next best measure: Videos removed by Policy Violation

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

Q3: **91,445,802** videos removed, Q4: **85,794,222** videos removed

GARM Category	Relevant Policy	Q4 2021	Q3 2021	Previous Period – Q2 2021	Commentary
Hate speech & acts of aggression	Hateful behavior	1.5%	1.5%	2.2%	Figure represents all Hateful Behavior, which includes <ul style="list-style-type: none"> • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hateful behavior - Slurs	1.5%	1.5%	2.2%	Figure represents all Hateful Behavior, which includes <ul style="list-style-type: none"> • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Illegal activities and regulated goods - Drugs, controlled substances, alcohol, and tobacco	19.5%	16.6%	20.8%	Figure represents all Illegal activities and regulated goods, which includes <ul style="list-style-type: none"> • Arms & Ammunition • Crime & Harmful acts to individuals and Society, Human Right Violations • Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol
Spam or Harmful Content	Integrity and authenticity - Spam and fake engagement	0.6%	0.5%	0.7%	Figure Represents all Integrity and Authenticity, including: <ul style="list-style-type: none"> • Online Piracy • Spam This figure also includes videos removed for Misinformation, as defined by TikTok Community Guidelines
Terrorism	Violent extremism	0.8%	0.9%	1%	
Debated Sensitive Social Issue	Hateful behavior	1.5%	1.5%	2.2%	Figure represents all Hateful Behavior, which includes <ul style="list-style-type: none"> • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity



Question 2: How safe is the platform for advertisers?

Not submitted

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors			<p>This is not something we currently track. However, content that appears either side of in-feed ads is moderated through AI and human reviewers.</p> <p>Because our ads are 100% share of voice (full screen), there is 0% on screen adjacency</p>
	Adult nudity and sexual activities			
Arms & Ammunition	Illegal activities and regulated goods - Weapons			
Crime & Harmful acts to individuals and Society, Human Right Violations	Illegal activities and regulated goods - Criminal Activities			
Death, Injury or Military Conflict	Violent and graphic content			
Online piracy	Integrity and authenticity - Intellectual property violations			
Hate speech & acts of aggression	Hateful behavior			
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hateful behavior – Slurs			
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Illegal activities and regulated goods - Drugs, controlled substances, alcohol, and tobacco			
Spam or Harmful Content	Integrity and authenticity - Spam and fake engagement			
Terrorism	Violent extremism			
Debated Sensitive Social Issue	Hateful behavior			



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

Next Best Measure: Automatic blocks of content

Percentage of violating videos removed within 24 hours

GARM Category	Relevant Policy	Q4 2021	Q3 2021	Previous Period – Q2 2021	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	96.2%	95.16%	95.4%	
	Adult nudity and sexual activities	90.5%	78.38%	90.0%	Minor Safety includes anything that may perpetuate the abuse, harm, endangerment, or exploitation of minors
Arms & Ammunition	Illegal activities and regulated goods - Weapons	95.7%	91.81%	95.7%	Figure represents all Illegal activities and regulated goods, which includes <ul style="list-style-type: none"> Arms & Ammunition Crime & Harmful acts to individuals and Society, Human Right Violations illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol
Crime & Harmful acts to individuals and Society, Human Right Violations	Illegal activities and regulated goods - Criminal Activities	95.7%	91.81%	95.7%	Figure represents all Illegal activities and regulated goods, which includes <ul style="list-style-type: none"> Arms & Ammunition Crime & Harmful acts to individuals and Society, Human Right Violations illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol
Death, Injury or Military Conflict	Violent and graphic content	94.9%	87.66%	94.3%	.
Online piracy	Integrity and authenticity - Intellectual property violations	75.7%	65.18%	86.2%	Figure Represents all Integrity and Authenticity, including: <ul style="list-style-type: none"> Online Piracy Spam This figure also includes videos removed for Misinformation, as defined by TikTok Community Guidelines



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

Next Best Measure: Automatic blocks of content

Percentage of violating videos removed within 24 hours

GARM Category	Relevant Policy	Q4 2021	Q3 2021	Previous Period – Q2 2021	Commentary
Hate speech & acts of aggression	Hateful behavior	82.5%	60.50%	80.8%	Figure represents all Hateful Behavior, which includes <ul style="list-style-type: none"> • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hateful behavior – Slurs	82.5%	60.50%	80.8%	Figure represents all Hateful Behavior, which includes <ul style="list-style-type: none"> • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Illegal activities and regulated goods - Drugs, controlled substances, alcohol, and tobacco	95.7%	91.81%	95.7%	Figure represents all Illegal activities and regulated goods, which includes <ul style="list-style-type: none"> • Arms & Ammunition • Crime & Harmful acts to individuals and Society, Human Right Violations • illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol
Spam or Harmful Content	Integrity and authenticity - Spam and fake engagement	75.7%	65.18%	86.2%	Figure Represents all Integrity and Authenticity, including: <ul style="list-style-type: none"> • Online Piracy • Spam This figure also includes videos removed for Misinformation, as defined by TikTok Community Guidelines
Terrorism	Violent extremism	89.5%	82.03%	90.1%	
Debated Sensitive Social Issue	Hateful behavior	82.5%	60.50%	80.8%	Figure represents all Hateful Behavior, which includes <ul style="list-style-type: none"> • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity



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

Next Best Measure: Automatic blocks of content

Percentage of violating videos removed within 24 hours

GARM Category	Relevant Policy	Q4 2021	Q3 2021	Previous Period – Q2 2021	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	98.2%	98.24%	97.6%	
	Adult nudity and sexual activities	90.3%	90.128%	90.3%	Minor Safety includes anything that may perpetuate the abuse, harm, endangerment, or exploitation of minors
Arms & Ammunition	Illegal activities and regulated goods - Weapons	96.8%	96.88%	97.1%	Figure represents all Illegal activities and regulated goods, which includes <ul style="list-style-type: none"> Arms & Ammunition Crime & Harmful acts to individuals and Society, Human Right Violations illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol
Crime & Harmful acts to individuals and Society, Human Right Violations	Illegal activities and regulated goods - Criminal Activities	96.8%	96.88%	97.1%	Figure represents all Illegal activities and regulated goods, which includes <ul style="list-style-type: none"> Arms & Ammunition Crime & Harmful acts to individuals and Society, Human Right Violations illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol
Death, Injury or Military Conflict	Violent and graphic content	96.1%	95.75%	94.9%	.
Online piracy	Integrity and authenticity - Intellectual property violations	85.5%	86%	88.3%	Figure Represents all Integrity and Authenticity, including: <ul style="list-style-type: none"> Online Piracy Spam This figure also includes videos removed for Misinformation, as defined by TikTok Community Guidelines



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

Next Best Measure: Automatic blocks of content

Percentage of violating videos removed within 24 hours

GARM Category	Relevant Policy	Q4 2021	Q3 2021	Previous Period – Q2 2021	Commentary
Hate speech & acts of aggression	Hateful behavior	76%	72.38%	72.9%	Figure represents all Hateful Behavior, which includes <ul style="list-style-type: none"> • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hateful behavior – Slurs	76%	72.38%	72.9%	Figure represents all Hateful Behavior, which includes <ul style="list-style-type: none"> • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Illegal activities and regulated goods - Drugs, controlled substances, alcohol, and tobacco	96.8%	96.88%	97.1%	Figure represents all Illegal activities and regulated goods, which includes <ul style="list-style-type: none"> • Arms & Ammunition • Crime & Harmful acts to individuals and Society, Human Right Violations • illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol
Spam or Harmful Content	Integrity and authenticity - Spam and fake engagement	85.5%	86%	88.3%	Figure Represents all Integrity and Authenticity, including: <ul style="list-style-type: none"> • Online Piracy • Spam This figure also includes videos removed for Misinformation, as defined by TikTok Community Guidelines
Terrorism	Violent extremism	92.2%	92.08%	89.4%	
Debated Sensitive Social Issue	Hateful behavior	76%	72.38%	72.9%	Figure represents all Hateful Behavior, which includes <ul style="list-style-type: none"> • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity



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

Next Best Measure: Percentage of violating videos removed by views

Percentage of violating videos removed before receiving a single view

GARM Category	Relevant Policy	Q4 2022	Q3 2022	Previous Period – Q2 2022	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	95.6%	95.16%	93.9%	<p>We continue to expand our system that detects and removes certain categories of violations at upload – including adult nudity and sexual activities, minor safety, and illegal activities and regulated goods. As a result, the volume of automated removals has increased, which improves the overall safety of our platform and enables our team to focus more time on reviewing contextual or nuanced content, such as hate speech, bullying and harassment, and misinformation.</p>
	Adult nudity and sexual activities	79.6%	78.38%	78.5%	
Arms & Ammunition	Illegal activities and regulated goods - Weapons	93%	91.81%	92.3%	
Crime & Harmful acts to individuals and Society, Human Right Violations	Illegal activities and regulated goods - Criminal Activities	93%	91.81%	92.3%	
Death, Injury or Military Conflict	Violent and graphic content	89.9%	87.66%	86.6%	
Online piracy	Integrity and authenticity - Intellectual property violations	67%	65.18%	67.9%	
Hate speech & acts of aggression	Hateful behavior	65.2%	60.50%	60.6%	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hateful behavior – Slurs	65.2%	60.50%	60.6%	
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Illegal activities and regulated goods - Drugs, controlled substances, alcohol, and tobacco	93%	91.81%	92.3%	
Spam or Harmful Content	Integrity and authenticity - Spam and fake engagement	67%	65.18%	67.9%	
Terrorism	Violent extremism	83.5%	82.03%	79.5%	
Debated Sensitive Social Issue	Hateful behavior	65.2%	60.50%	60.6%	



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

Authorized Metric: Removal of Violating Accounts

GARM Category	Relevant Policy	Q4 2021	Q3 2021	Previous Period – Q2 2021	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	We removed 24,107,316 Accounts for violating community guidelines	We removed 17,005,726 Accounts for violating community guidelines	We removed 14,871,412 Accounts for violating community guidelines	Account Removals represents figure across all community guidelines\
	Adult nudity and sexual activities				
Arms & Ammunition	Illegal activities and regulated goods - Weapons				
Crime & Harmful acts to individuals and Society, Human Right Violations	Illegal activities and regulated goods - Criminal Activities				
Death, Injury or Military Conflict	Violent and graphic content				
Online piracy	Integrity and authenticity - Intellectual property violations				
Hate speech & acts of aggression	Hateful behavior				
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hateful behavior – Slurs				
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Illegal activities and regulated goods - Drugs, controlled substances, alcohol, and tobacco				
Spam or Harmful Content	Integrity and authenticity - Spam and fake engagement				
Terrorism	Violent extremism				
Debated Sensitive Social Issue	Hateful behavior				



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

Authorized Metric: Appeals & Reinstatements

Content removed by TikTok and then appealed by users

GARM Category	Relevant Policy	Q4 2021	Q3 2021	Previous Period – Q2 2021	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	We reinstated 4,727,382 videos after they were appealed	We reinstated 5,535,378 videos after they were appealed	We reinstated 4,663,387 videos after they were appealed	Content reinstatement represents figure across all community guidelines\
	Adult nudity and sexual activities				
Arms & Ammunition	Illegal activities and regulated goods - Weapons				
Crime & Harmful acts to individuals and Society, Human Right Violations	Illegal activities and regulated goods - Criminal Activities				
Death, Injury or Military Conflict	Violent and graphic content				
Online piracy	Integrity and authenticity - Intellectual property violations				
Hate speech & acts of aggression	Hateful behavior				
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hateful behavior – Slurs				
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Illegal activities and regulated goods - Drugs, controlled substances, alcohol, and tobacco				
Spam or Harmful Content	Integrity and authenticity - Spam and fake engagement				
Terrorism	Violent extremism				
Debated Sensitive Social Issue	Hateful behavior				

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.

Our latest transparency report includes data from Q3 2021 (July–September 2021) and Q4 2021 (October–December 2021). During this reporting period, we continued to support the health and wellbeing of our community with innovative features like [hair pattern search](#), which empowers users to search for hair inspiration across hair types. We also [updated our ad policies](#) to prohibit all ads with weight loss language and imagery as part of our efforts to create a culture where healthy living and healthy body image can thrive.

In June 2021, we received brand safety certification from the Trustworthy Accountability Group (TAG), a global certification body that aims to fight criminal activity and protect brand safety in digital advertising. Thanks to the broad reach of the industry’s largest brand safety certification program, our TAG status extends to APAC, Europe, Latin America and North America.

In October 2021, the Media Rating Council (MRC), an independent non-profit organization that sets standards for digital advertising, [accredited Pinterest](#) for two key metrics: display Pin impressions and display Pin clicks. Obtaining accreditation means that Pinterest met or exceeded compliance with industry standards to measure display impressions and display clicks. This includes invalid traffic filtration for activity from bots and crawlers. So, advertisers can find comfort knowing that Pinterest is working hard to show only real actions.

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 Q3 and Q4 2021 and was collected on March 30, 2022.



Question 1: How safe is the platform for consumers?

Next Best Measure: Reach¹ of Pins deactivated for violating policy

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

Pinterest Policy ²	Latest Period								Previous Period							
	Q4 2021				Q3 2021				Q2 2021				Q1 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	88%	10%	2%	0.6%	82%	14%	2%	1%	81%	15%	3%	1.3%	80%	15%	3%	1.6%
Adult sexual services	89%	9%	2%	0.6%	0.4% ³	47%	29%	24%	0.7% ⁴	45%	32%	22%	1.5% ⁵	41%	28.1%	30%
Civic misinformation	99.5%	0.4%	0.06%	0.04%	10%	78%	7%	5%	70%	26%	3%	0.8%	80%	16%	3%	0.9%
Conspiracy theories	89%	8%	1%	2%	93%	6%	0.3%	0.1%	95%	5%	0.3%	0.1%	92%	6%	1%	0.3%
Dangerous goods and activities	76%	16%	5%	3%	95%	4%	0.9%	0.3%	97%	2%	0.2%	0.1%	95%	4%	0.6%	0.4%
Graphic violence and threats	59%	13%	10%	19%	52%	29%	8%	11%	26%	28%	21%	26%	96%	3%	0.4%	0.5%
Harassment and criticism	69%	26%	3%	2%	86%	12%	2%	1%	79%	18%	3%	1%	87%	10%	2%	1%
Hateful activities	97%	0.8%	0.7%	1%	56%	19%	11%	14%	22%	34%	18%	26%	70%	13%	7%	11%
Medical misinformation	76%	9%	5%	9%	75%	20%	4%	1%	65%	19%	8%	8%	74%	16%	3%	8%
Self-injury and harmful behavior	97%	2%	0.3%	0.3%	83%	12%	3%	2%	88%	8%	2%	2%	86%	12%	2%	1%
Spam	75%	19%	4%	1%	63%	24%	9%	4%	61%	20%	12%	7%	65%	15%	13%	7%

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 and Harassment and criticism in Q2 do 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 Of the 1,263 Pins deactivated in Q2 2021 for violating our Adult sexual services policy, 984 were seen by fewer than 100 users in that reporting period.

⁵ Of the 533 Pins deactivated in Q1 2021 for violating our Adult sexual services policy, 375 were seen by fewer than 100 users in that reporting period.



Question 1: How safe is the platform for consumers?

Next Best Measure: Reach¹ of Pins deactivated for violating policy

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

Pinterest Policy ²	Latest Period								Previous Period							
	Q4 2021				Q3 2021				Q2 2021				Q1 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	88%	10%	2%	0.6%	82%	14%	2%	1%	81%	15%	3%	1.3%	80%	15%	3%	1.6%
Adult sexual services	89%	9%	2%	0.6%	0.4% ³	47%	29%	24%	0.7% ⁴	45%	32%	22%	1.5% ⁵	41%	28.1%	30%
Civic misinformation	99.5%	0.4%	0.06%	0.04%	10%	78%	7%	5%	70%	26%	3%	0.8%	80%	16%	3%	0.9%
Conspiracy theories	89%	8%	1%	2%	93%	6%	0.3%	0.1%	95%	5%	0.3%	0.1%	92%	6%	1%	0.3%
Dangerous goods and activities	76%	16%	5%	3%	95%	4%	0.9%	0.3%	97%	2%	0.2%	0.1%	95%	4%	0.6%	0.4%
Graphic violence and threats	59%	13%	10%	19%	52%	29%	8%	11%	26%	28%	21%	26%	96%	3%	0.4%	0.5%
Harassment and criticism	69%	26%	3%	2%	86%	12%	2%	1%	79%	18%	3%	1%	87%	10%	2%	1%
Hateful activities	97%	0.8%	0.7%	1%	56%	19%	11%	14%	22%	34%	18%	26%	70%	13%	7%	11%
Medical misinformation	76%	9%	5%	9%	75%	20%	4%	1%	65%	19%	8%	8%	74%	16%	3%	8%
Self-injury and harmful behavior	97%	2%	0.3%	0.3%	83%	12%	3%	2%	88%	8%	2%	2%	86%	12%	2%	1%
Spam	75%	19%	4%	1%	63%	24%	9%	4%	61%	20%	12%	7%	65%	15%	13%	7%

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 and Harassment and criticism in Q2 do 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 Of the 1,263 Pins deactivated in Q2 2021 for violating our Adult sexual services policy, 984 were seen by fewer than 100 users in that reporting period.

⁵ Of the 533 Pins deactivated in Q1 2021 for violating our Adult sexual services policy, 375 were seen by fewer than 100 users in that reporting period.



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

Authorized Metric: Reach¹ of Pins deactivated for violating policy

Pinterest Policy ²	Latest Period								Previous Period							
	Q4 2021				Q3 2021				Q2 2021				Q1 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	88%	10%	2%	0.6%	82%	14%	2%	1%	81%	15%	3%	1.3%	80%	15%	3%	1.6%
Adult sexual services	89%	9%	2%	0.6%	0.4% ³	47%	29%	24%	0.7% ⁴	45%	32%	22%	1.5% ⁵	41%	28.1%	30%
Civic misinformation	99.5%	0.4%	0.06%	0.04%	10%	78%	7%	5%	70%	26%	3%	0.8%	80%	16%	3%	0.9%
Conspiracy theories	89%	8%	1%	2%	93%	6%	0.3%	0.1%	95%	5%	0.3%	0.1%	92%	6%	1%	0.3%
Dangerous goods and activities	76%	16%	5%	3%	95%	4%	0.9%	0.3%	97%	2%	0.2%	0.1%	95%	4%	0.6%	0.4%
Graphic violence and threats	59%	13%	10%	19%	52%	29%	8%	11%	26%	28%	21%	26%	96%	3%	0.4%	0.5%
Harassment and criticism	69%	26%	3%	2%	86%	12%	2%	1%	79%	18%	3%	1%	87%	10%	2%	1%
Hateful activities	97%	0.8%	0.7%	1%	56%	19%	11%	14%	22%	34%	18%	26%	70%	13%	7%	11%
Medical misinformation	76%	9%	5%	9%	75%	20%	4%	1%	65%	19%	8%	8%	74%	16%	3%	8%
Self-injury and harmful behavior	97%	2%	0.3%	0.3%	83%	12%	3%	2%	88%	8%	2%	2%	86%	12%	2%	1%
Spam	75%	19%	4%	1%	63%	24%	9%	4%	61%	20%	12%	7%	65%	15%	13%	7%

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 and Harassment and criticism in Q2 do 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 Of the 1,263 Pins deactivated in Q2 2021 for violating our Adult sexual services policy, 984 were seen by fewer than 100 users in that reporting period.

⁵ Of the 533 Pins deactivated in Q1 2021 for violating our Adult sexual services policy, 375 were seen by fewer than 100 users in that reporting period.



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

Authorized Metric: Distinct images deactivated¹, Pins deactivated², Boards deactivated³, Accounts deactivated⁴

Violating content deactivated by Pinterest

Pinterest Policy	Latest Period								Previous Period ⁵							
	Q4 2021				Q3 2021				Q2 2021				Q1 2021			
	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	1,108,006	30,695,545	44,155	6,769	1,279,861	42,968,305	61,114	8,653	1,486,098	46,130,733	131,664	13,468	2,029,934	47,120,204	179,994	17,659
Adult sexual services	1,610	26,703	120	121	222	225	151	152	1,245	1,263	708	500	520	533	348	271
Civic misinformation	779	48,741	23	0	743	889	92	8	349	2,316	92	9	2,044	6,895	546	45
Conspiracy theories	2,411	4,507	237	11	8,374	536,455	429	39	16,204	1,148,947	451	81	24,134	166,189	592	116
Dangerous goods and activities	14,736	31,186	591	177	44,277	1,017,444	712	223	938,363	12,286,669	1,814	501	108,534	839,389	2,344	403
Graphic violence and threats	3,435	6,085	769	89	4,026	18,501	1,085	106	4,005	7,173	767	72	3,376	254,455	1,184	243
Harassment and criticism	4,318	210,880	803	151	4,257	244,688	851	271	7,238	1,238,782	1,025	292	5,540	124,713	977	594
Hateful activities	3,848	107,295	409	36	3,482	9,805	1,758	107	3,418	6,086	739	126	2,487	8,823	665	82
Medical misinformation	3,564	6,370	491	14	4,237	138,491	383	10	4,869	19,852	462	34	4,256	11,097	473	35
Self-injury and harmful behavior	16,761	395,682	82,518	1,047	4,156	85,328	575	37	3,824	81,444	828	123	2,945	302,918	545	79
Spam	61,834	121,639	0	1,736,742	81,753	233,303	0	2,435,683	118,054	378,490	0	3,104,409	336,821	1,042,960	1	4,420,965

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 Redundant logging in Q1 and Q2 2021 resulted in the reported numbers reflecting a slight overcounting of board and account deactivations for that period. As a result, their respective deactivation numbers in the first half of 2021 are not directly comparable to those in the second half.



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

Authorized Metric: Distinct images deactivated¹, Pins deactivated², Boards deactivated³, Accounts deactivated⁴

Violating content deactivated by Pinterest

Pinterest policy	Commentary
Adult content	In Q1 2021, we launched new machine learning tools to detect and deactivate boards for violating our adult content policy. As a result, there was an increase in the number of boards and accounts deactivated. Our content policies and moderation practices are always evolving to better keep our platform safe for all users. So far, we've primarily focused our use of these new machine learning tools on enforcing against our adult content policy, and we're currently exploring how they may work for other types of policy violations.
Adult sexual services	In Q4 2021, we made changes to the tools we use to deactivate this type of content. As a result of these improvements, content that previously would have been deactivated under our adult content policy is now deactivated under our adult sexual services policy. The increase in the portion of Pins deactivated with hybrid tools, as well as the increase in Pins deactivated, reflect these changes.
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 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 <i>actually</i> violates this policy before the Pin was deactivated.
Conspiracy theories	In Q2 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. There was a relative decrease in the number of Pins deactivated for violating this policy in Q3 and Q4 2021.
Dangerous goods and activities	In Q1 and Q2 2021, we performed a sweeping cleanup across the platform for content violating our dangerous goods and activities policy. This effort, applied to organic Pins and not promoted ads or products, generated a temporary spike in Pins deactivated. This rise is not due to any known increase of violative content. There was a relative decrease in the number of Pins deactivated for violating this policy in Q3 and Q4 2021.
Graphic violence and threats	There was a relative increase in the amount of content deactivated for violating our graphic violence and threats policy during the US presidential transition in Q1 2021. In Q3 2021, we saw relatively fewer Pins deactivated compared to Q4 2021. Of the Pins we deactivated in Q3 2021, 89% 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.



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

Authorized Metric: Distinct images deactivated¹, Pins deactivated², Boards deactivated³, Accounts deactivated⁴

Violating content deactivated by Pinterest

Pinterest policy	Commentary
Harassment and criticism	<p>We determined that a small handful of the distinct images deactivated in Q2 2021, and the more than 990,000 Pins deactivated as machine-identified matches, were incorrectly deactivated, and we reinstated that content after spotting the error. Of the Pins that we believe were correctly deactivated, 79% were never seen by users in that reporting period.</p> <p>We've included those false positives in the Q2 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 <i>actually</i> violates this policy before the Pin was deactivated.</p>
Hateful activities	<p>In Q1 and Q2 2021, we saw a relative decrease in boards and accounts deactivated compared to Q4 2020. We deactivated fewer Pins in Q2 2021, and 74% of those Pins were seen by fewer than 100 users in that reporting period.</p> <p>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.</p>
Medical misinformation	<p>As COVID-19 vaccines continue to be made available around the world, we removed content that violated our medical misinformation policy and worked with health experts throughout the year to discuss issues within the community. We saw a relative increase in the number of Pins deactivated in Q3 2021, and 99% of these Pins were seen by fewer than 100 users in this reporting period.</p>
Self-injury and harmful behavior	<p>We've continued to invest in work improving our policies around self-harm content and providing compassionate support for those in need. Of the Pins deactivated in Q1 2021 for violating our self-injury and harmful behavior policy, 86% were never seen by users in that reporting period. In Q4 2021, we performed a sweeping cleanup across the platform for content violating this policy that generated a temporary spike in Pins deactivated. This rise is not due to any known increase of violative content. More importantly, 97% of those Pins were never seen by users in that reporting period.</p>
Spam	<p>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.</p>

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

Pinterest Policy	Latest Period						Previous Period					
	Q4 2021			Q3 2021			Q2 2021			Q1 2021		
	Automated ¹	Manual ²	Hybrid ³	Automated	Manual	Hybrid	Automated	Manual	Hybrid	Automated	Manual	Hybrid
Adult content	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%	0%	<1%	>99%
Adult sexual services	0%	<1%	>99%	0%	100%	0%	0%	100%	0%	0%	100%	0%
Civic misinformation	0%	<1%	>99%	0%	87%	13%	0%	16%	84%	0%	96%	4%
Conspiracy theories	0%	90%	10%	0%	<1%	>99%	0%	<1%	>99%	0%	2%	98%
Dangerous goods and activities	14%	41%	45%	<1%	2%	98%	<1%	<1%	>99%	<1%	5%	95%
Graphic violence and threats	0%	62%	38%	0%	48%	53%	0%	79%	21%	0%	2%	98%
Harassment and criticism	0%	2%	98%	0%	1%	99%	0%	<1%	>99%	0%	11%	89%
Hateful activities	0%	5%	95%	0%	46%	54%	0%	60%	40%	0%	30%	70%
Medical misinformation	72%	28%	<1%	2%	25%	74%	31%	30%	39%	48%	12%	40%
Self-injury and harmful behavior	0%	1%	99%	0%	5%	95%	0%	10%	90%	0%	1%	99%
Spam	>99%	<1%	0%	>99%	<1%	0%	>99%	<1%	0%	>99%	<1%	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

Pinterest Policy	Latest Period				Previous Period			
	Q4 2021		Q3 2021		Q2 2021		Q1 2021	
	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated	Accounts appealed	Accounts reinstated
Adult content	1,495	774	1,923	1,119	2,398	1,442	2,335	1,634
Adult sexual services	16	1	13	0	14	4	11	3
Civic misinformation	1	1	3	3	4	3	13	11
Conspiracy theories	13	2	19	4	13	5	20	11
Dangerous goods and activities	4	1	14	3	17	5	14	4
Graphic violence and threats	19	3	33	18	29	13	15	10
Harassment and criticism	28	15	22	13	33	22	57	43
Hateful activities	10	5	35	21	23	12	22	13
Medical misinformation	4	3	2	0	11	6	7	5
Self-injury and harmful behavior	166	133	10	9	28	25	13	12
Spam	103,257	79,054	101,832	77,936	80,624	61,050	112,139	84,541

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: Violative 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	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	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			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: Violative 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: Violative 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	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	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			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: Violative 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 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
		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				



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						Snap does not currently offer an appeals process, and therefore does not 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						

Twitch

To ensure that Twitch is a place where everyone can connect and create together, we believe that creators and their communities must feel safe on our service. Our goal is to foster a community that supports, grows, and sustains creators; provides a welcoming and entertaining environment for viewers; and prevents and eliminates illegal, negative, and harmful interactions.

At Twitch, we believe everyone in our community – creators, viewers, moderators, and our employees – plays a big role in promoting the health and safety of our community. Through our Community Guidelines, we make clear what expressions and behaviors are allowed on the service, and what are not. At the platform-level, we leverage machine detection to scan content and flag it for review, user reports submitted by our users to report content that violates our guidelines, and review and enforcement handled by a group of internal highly trained and experienced content moderators. Creators and moderators (colloquially known as “mods”) also use tools that we provide, such as AutoMod and Mod View, to enforce Twitch service wide standards or to set higher standards in their own channels. The result is a layered approach to safety – one that combines Twitch service level efforts (through proactive technology and staffing) channel level tooling for creators and viewer level tools all working together.

2021 H2 Overview

We’re constantly investing in tooling to ensure all of our users are authentically, safely, and meaningfully engaging with each other. In H2 2021, we took considerable steps to further improve safety on our service. We launched two powerful new tools to fight chat-based harassment: Suspicious User Detection (machine learning that flags potential channel-ban evaders) and Phone-Verified Chat (where users need to have a verified phone number to participate in live chats). We also banned linked accounts when they participated in targeted attacks or spam distribution and created new educational content to help creators understand what steps they can take to protect themselves with our built-in moderation tools. Product updates also allowed us to take concerted action against bot-related accounts which resulted in an 86% decrease in spam or harmful content being shared on Twitch.

Methodology

Our Community Guidelines around violative content on Twitch covers 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 three categories for this report – Online Piracy, Arms and Ammunition, and Debated Sensitive Issues. We are working on determining relevant subjects to these categories, and establishing the underlying technology and data to measure these areas.

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.

Twitch

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 a complete breakdown of 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	0.05%	0.06%	We prohibit 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	0.01%	0.02%	We prohibit streamers from being 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	
Illegal Drugs / Tobacco / cigarettes / Vaping / Alcohol	<0.01%	<0.01%	We prohibit 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 prohibit 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	0.12%	0.03%	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%	0.02%	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	0.14%	0.19%	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 a complete breakdown of 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	0.22%	0.28%	We prohibit 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%	0.04%	We prohibit streamers from being 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	
Illegal Drugs / Tobacco / ecigarettes / Vaping / Alcohol	<0.01%	<0.01%	We prohibit 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	0.01%	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 prohibit 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	0.04%	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 .



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	27,920	18,843	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. Although we saw an increase in enforcement actions in H2, our percent of advertising impressions remained low on violative content and decreased by 50% from H1 2021. Additionally, our % of total HWs remained consistent from the previous reporting period.
Crime and Harmful Acts to Individuals and Society, Human Right Violations	114,344	131,168	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 a complete breakdown of our hate and harassment reports and enforcement, please see our Transparency Report.
Death, Injury or Military Conflict	3,932	5,980	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	102,682	130,544	Following the launch of our new tools to help Creators and Mods protect their communities, namely Suspicious User Detection and Phone-Verified Chat, we saw a significant decrease in enforcements. We also created additional educational content to boost awareness of our suite of built-in moderation tools.
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic, or repulsive content	31,661	24,353	We prohibit streamers from being fully or partially nude. Additionally, content that exclusively focuses on extreme or gratuitous gore and violence is prohibited.



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	
Illegal Drugs / Tobacco / E-cigarettes / Vaping / Alcohol	18	20	We prohibit 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	2,151,932	15,719,228	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.
Terrorism	34	55	We prohibit 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	254,148	739,168	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 4: How does the platform perform at correcting mistakes

Authorized Metric: Total Enforcement Actions

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

GARM Category	Latest Period		Previous Period	Commentary
	Appeal Rate	Reinstatement Rate	Appeal and Reinstatement Rate	
Adult & Explicit Sexual Content	1.76%	2.68%	0.28% - 1.33%	This reporting period, we launched powerful moderation tools that caused Spam-related enforcement actions to decrease by 13 million. Due to the decrease in overall enforcement actions, our relative appeals rate increased; however, we did not see a large increase in the total number of appeals.
Arms & Ammunition				
Crime and Harmful Acts to Individuals and Society, Human Right Violations				
Death, Injury or Military Conflict				
Online piracy				
Hate speech and acts of aggression				
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust				
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol				
Spam or Harmful Content				
Terrorism				
Debated Sensitive Social Issue				

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 Group 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?

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

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

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

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

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









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

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



Question	Authorized Metric								
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 3: Date ranges for platform data submitted

	Q3 2020	Q4 2020	Q1 2021	Q2 2021	Q3 2021	Q4 2021
			PREVIOUS PERIOD		LATEST PERIOD	
			PREVIOUS PERIOD		LATEST PERIOD	
			PREVIOUS PERIOD		LATEST PERIOD	
	PREVIOUS PERIOD		LATEST PERIOD			
				PREVIOUS PERIOD	LATEST PERIOD	
			PREVIOUS PERIOD		LATEST PERIOD	
			PREVIOUS PERIOD		LATEST PERIOD	
			PREVIOUS PERIOD		LATEST PERIOD	

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

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

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

Level 1: Brand Safety Controls & Measurement

Level 2: Brand Safety Integrations

Level 3: Brand Safety Transparency Reporting

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

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

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

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

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

GARM Aggregated Measurement Report

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



World Federation of Advertisers

London, Brussels, Singapore, New York

wfanet.org

info@wfanet.org

+32 2 502 57 40

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

youtube.com/wfamarketers

linkedin.com/company/wfa