GARM Global Alliance for
Responsible Media

GARM Aggregated Measurement Report

Volume 1 | April 2021

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 throughthe 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 policy 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:

- Assessing Safety to Inform Media Selection considerations related to content safety.
- Assessing Progress on Safety Enforcement
- Assessing Topical Exposure and/ or Progress
- Determining 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 to 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 evolve undoubtedly over time. We foresee the evolution of the report coming via the following ways:

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

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

We're excited to launch this report with the partnership and collaboration within GARM, notably with YouTube, Facebook, Instagram, Twitter, TikTok, Snap, and Pinterest. Due to time constraints, Twitch (who joined GARM in March 2021), is not included in our inaugural report; however, we're looking forward to including them in our next Aggregated Measurement Report, later this year (information on Twitch practices can be found in their first ever Transparency Report, released in March 2021).

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

Executive Summary

We're excited to share GARM's inaugural volume of the Aggregated Measurement Report. This is the result of 9-months of collaborative workshops to advance the existing first-party transparency reporting available today. In our uncommon collaboration, we've created a report creating an unprecedented level of visibility across the industry in how efforts to remove harmful content from ad-supported digital media are progressing.

The first area of value creation for the Aggregated Measurement Report is around access, focus and rigor. The report brings data into one place, with key data in a common framework. This increased rigor and structure will help advertisers, agencies, and platforms.

The second area of value creation from the report is in observations and learnings. We're proceeding cautiously, because this is the first volume of the report and because we're looking at high-level data from the most recent 12 months of platform reporting. That said, we're starting to draw initial learnings. We recognize that there's more historical data to mine and there's future data to analyze.

Before we look across the industry to form learnings, we need to recognize the context of the latest period of data provided. While there are some differences in the time series for each of the platforms, there were environmental factors in 2020 that had a massive impact on brand safety activity for every platform: the ongoing pandemic and elections (in 40+ markets, not just the US). As such, it is fair to recognize that the landscape was marked with concerns about content-based misinformation, and the operations of platform and brand safety were initially challenged by working remotely. Suffice it to say that 2020 was an exceptional year for content safety themes and platform safety operations.

GARM is committed to producing this report every 6 months and continuing to track trends and identify areas for collaboration to prevent harmful content from being monetized.

Learning 1: Numbers of content removals remain consistent, with Spam, Sexual Content and Hate Speech as the biggest enforcement areas.

Content removal has remained roughly consistent in the two periods analyzed at 5.3B+ pieces of content removed, down slightly from the time period prior. But looking at both time periods, more than four out of five pieces of content removed stem from three GARM content categories – Spam, Sexual Content, and Hate Speech and Acts of Aggression.

Learning 2: Hate Speech and Acts of Aggression is an Area of Intensified Enforcement

Hate Speech and Acts of Aggression have been the focus of industry attention, and we are seeing intensified enforcement across metrics being shared. Our first point of reference is seeing the advances made by YouTube in this area specific to account removals. Our second point of reference is Facebook's reduction in prevalence for hate speech, which decreased by 20% from Q3 to Q4.

Learning 3: Automation and Machine Learning Accelerated During the Pandemic

Pandemic-related workplace restrictions had an impact on how platforms approached content moderation; automated and machine learning methods for content assessment and removal grew in its importance consistently across platforms. Facebook, Instagram, and Pinterest report on the role for machine blocking at a category level, showing highest machine moderation in areas such as Terrorism, Violent Graphic Content, Crime and Harmful Acts, and Arms and Ammunition.

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 discoverability of content that brushes right up against our policy line and reward trusted partners.**

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 or hate speech. These guidelines apply to all types of content on the platform, including videos, comments, links, and thumbnails. Over the past several years, machine learning has transformed our ability to tackle how we remove violative content at scale. From July 2018 to December 2020, we removed over 83 million videos and 7 billion comments for violating our Community Guidelines. Because of our investments in machine learning, as of Q4 2020 we are now able to detect 94% of all violative content on YouTube by automated flagging – and 75% of flagged content gets removed with fewer than 10 views.

We have also made huge investments in other areas of critical importance, like transparency. Several years ago, we 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. 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 our last four transparency reports into this resource.

We also enforce a second set of policies, our **Ad Friendly Guidelines,** which set the standard for which videos are eligible for ads. These guidelines are more restrictive than our Community Guidelines and adhere to the GARM brand safety floor. In some cases, our Ad Friendly Guidelines may be even more restrictive. Data regarding Ad Friendly Guidelines enforcement is not currently included in our Transparency Report."

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

Methodology for Metrics

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

YouTube

The YouTube Community Guidelines enforcement report contains data on actions YouTube takes with regard to content on the platform that violates our policies. This includes:

- Flagging (human and automated)
- Video, channel, and comment removals
- Appeals and reinstatements
- Highlighted policy verticals

The report first launched in April 2018. Since then, 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 this resource, we have aggregated data from 2020 and, where possible, aggregated that data on a bi-annual basis.

January through June 2020

Between January and June 2020, YouTube removed over 17.5 million videos for violating Community Guidelines. The vast majority (>94%) of these videos were first flagged by machines rather than humans. YouTube terminated over 3.9 million channels for violating our Community Guidelines, the overwhelming majority which were terminated for violating our spam policies. Our violative view rate ranged from 0.17-0.20% in Q1 2020 to 0.18-0.21% in Q2 2020. This means that out of every 10,000 views on YouTube, only 17-20 came from violative content in Q1, and only 18-21 came from violative content in Q2. YouTube removed more than 2.8 billion comments, the majority of which were spam; 99% of removed comments were detected automatically. Just over 491k video removals were appealed, and we reinstated ~201k of those videos.

We normally rely on a combination of people and technology to enforce our policies. Machine learning helps detect potentially harmful content, and then sends it to human reviewers for assessment. Human review is not only necessary to train our machine learning systems, it also serves as a check, providing feedback that improves the accuracy of our systems over time. Each quarter, millions of videos that are first flagged by our automated systems are later evaluated by our human review team and determined not to violate our policies. Numbers fluctuate every quarter due to a number of factors, including changes to our policies and the evolution of our enforcement.

Given the unprecedented circumstances created by the COVID-19 pandemic, in March of 2020, we took steps to protect our employees and extended workforce during the COVID-19 pandemic. One major step was to rely more on technology to quickly identify and remove content that violates our Community Guidelines so that our teams that review content could safely remain at home. The second quarter of 2020 was the first full quarter we operated under this modified enforcement structure. Because of choices we made to prioritize the safety of the community, we removed the most videos (11.4M) we've ever removed in a single quarter from YouTube. This also resulted in higher appeals and reinstatements. Child safety in particular is a priority area for us when it comes to user safety. During this time, it was no exception and as a result was a policy area where we cast a much wider net for removals as you'll see reflected in the appendix.

July through December 2020

Between July and December 2020, YouTube removed over 17 million videos for violating our Community Guidelines. The vast majority (>90%) of these videos were first flagged by machines rather than humans. YouTube terminated over 3.8 million channels for violating our Community Guidelines, the overwhelming majority which were terminated for violating our spam policies. Our violative view rate ranged from 0.15-0.17% in Q3 2020 to 0.16-.018% in Q4 2020. This means that out of every 10,000 views on YouTube, only 15-17 came from violative content in Q3, and only 16-18 came from violative content in Q4. YouTube removed more than 2 billion comments, the majority of which were spam; 99% of removed comments were detected automatically. Just over 432k video removals were appealed, and we reinstated ~165k of those.

We also updated several of our policies to protect our community in response to evolving consumer trends. For example, in October we expanded both our hate and harassment policies to prohibit content that targets an individual or group with conspiracy theories that have been used to justify real-world violence. One example would be content that threatens or harasses someone by suggesting they are complicit in one of these harmful conspiracies, such as QAnon or Pizzagate. In these areas of Harassment, Harmful or dangerous, and Hateful content, in the second half of 2020, we removed more channels (actors) than in previous periods to reflect these policy updates. (Additional details in the appendix.)

Our Community Guidelines prohibit spam, scams, or other manipulated media, coordinated influence operations, and any content that seeks to incite violence. Since September 2020, we've terminated over 8,000 channels and thousands of harmful and misleading elections-related videos for violating our existing policies. Over 77% of those removed videos were taken down before they had 100 views.

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.

Question 1: How safe is the platform for consumers?

Authorized Metric: Violative View Rate

YouTube measures consumer safety as the percentage of removed videos by views and the percentage of views as first detected by machines

	Latest Period		Previous Period	
GARM Metric	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**
Violative View Rate	0.15-1.17%	0.16-0.18%	0.17-0.20%	0.18-0.21%

Question 2: How safe is the platform for advertisers?

Authorized Metric: Advertising Safety Error Rate

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

	Latest Period		Previous Period	
GARM Metric	Q3 2020	Q4 2020	Q1 2020	Q2 2020
Advertising Safety Error Rate	<1%	<1%	<1%	<1%

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

Authorized Metric: Content Removed, Actors Removed, Comments Actioned

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 govern content that can live on YouTube

– Enforcement of these guidelines is reflected in our quarterly Community Guidelines Enforcement Report

YouTube Policy	Content Removed[1]		Actors Removed[2]		Comments Removed	
	Latest Period Q3&Q42020	Previous Period Q1&Q22020	Latest Period Q3&Q42020	Previous Period Q1&Q22020	Latest Period Q3&Q42020	Previous Period Q1&Q22020
Nudity or sexual	3,046,206	2,533,712	252,878	268,725	2,069,365	491,557
Child safety	6,321,184	5,302,746	69,917	67,170	568,547,152	293,534,691
Harmful or dangerous	455,924	509,581	5,231	207	813,921	950,911
Promotion of violence and violent extremism	273,475	1,180,691	22,880	14,025	885,875	228,478
Harassment and cyberbullying	121,337	144,183	79,194	21,147	333,553,579	739,280,524
Violent or graphic	3,037,725	1,908,720	574	330	23,788	27,578
Spam, misleading and scams	3,446,883	5,487,458	3,129,658	3,566,691	1,046,763,757	1,596,097,615
Hateful or abusive	182,496	187,207	224,929	5,980	93,275,748	195,043,068
Impersonation	N/A	N/A	62,580	10,559	N/A	N/A
Other	309,402	258,406	546	4,487	541,862	292,914

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

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

Authorized Metric: 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

	Latest Period		Previous Period	
YouTube Metric	Q3 2020	Q4 2020	Q1 2020	Q2 2020
Total Video Removals	7,872,684	9,321,948	6,111,008	11,401,696
Removed videos by views: 0 views	42.6%	35.9%	49.9%	42.0%
Removed videos by views: 1-10	33.8%	35.9%	27.4%	33.7%
Removed videos by views: 10+	23.6%	28.2%	22.7%	24.3%
Removed videos by views as detected by machines: 0 Views	45.2%	37.6%	53.0%	44.0%
Removed videos by views as detected by machines: 1-10	35%	37.5%	28.1%	34.8%
Removed videos by views as detected by machines: 10+	19.8%	24.9%	18.9%	21.2%

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

Authorized Metric: Appeals, Reinstatements

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

	Latest Period	Previous Period
GARM Metric	**Q3 & Q4 2020**	**Q1 & Q2 2020**
Content Appealed: Videos	432,934	491,380
Content Reinstated: Video	165,490	201,680

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

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

GARM Brand Safety Floor Category + Definition • Defines content that can monetize. • Aligned with YouTube's Ad Friendly Guidelines, a higher bar than Community Guidelines.	Relevant YouTube Community Guidelines – Governs content that can live on YouTube. – Our Community Guidelines Enforcement Report measures our enforcement of these guidelines.
Adult & Explicit Sexual Content • Illegal sale, distribution, and consumption of child pornography • Explicit or gratuitous depiction of sexual acts, and/or display of genitals, real or animated	Nudity and Sexual Content Explicit content meant to be sexually gratifying is not allowed on YouTube. Posting pornography may result in content removal or channel termination. Videos containing fetish content will be removed or age-restricted. In most cases, violent, graphic, or humiliating fetishes are not allowed on YouTube.
	Child Safety YouTube doesn't allow content that endangers the emotional and physical well-being of minors. A minor is defined as someone under the legal age of majority -- usually anyone younger than 18 years old in most countries/regions.
Arms & Ammunition • Promotion and advocacy of Sales of illegal arms, rifles, and handguns • Instructive content on how to obtain, make, distribute, or use illegal arms • Glamorization of illegal arms for the purpose of harm to others • Use of illegal arms in unregulated environments	Firearms Content intended to sell firearms, instruct viewers on how to make firearms, ammunition, and certain accessories, or instruct viewers on how to install those accessories is not allowed on YouTube. YouTube shouldn't be used as a platform to sell firearms or accessories noted below. YouTube also doesn't allow live streams that show someone holding, handling, or transporting a firearm.
Crime & Harmful acts to individuals and Society, Human Right Violations • Graphic promotion, advocacy, and depiction of willful harm and actual unlawful criminal activity - Explicit violations/demeaning offenses of Human Rights (e.g. human trafficking, slavery, self-harm, animal cruelty etc.) • Harassment of bullying of individuals and groups	Harmful or Dangerous Content YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death.
	Hate Speech Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status
	Harassment and cyberbullying Content that threatens individuals is not allowed on YouTube. We also don't allow content that targets an individual with prolonged or malicious insults based on intrinsic attributes. These attributes include their protected group status or physical traits.
Death, Injury or Military Conflict • Promotion, incitement or advocacy of violence, death or injury • Murder or willful bodily harm to others • Graphic depictions of willful harm to others • Incendiary content provoking, enticing, or evoking military aggression • Live action footage/photos of military actions & genocide or other war crimes	Violent or Graphic Content Violent or gory content intended to shock or disgust viewers, or content encouraging others to commit violent acts are not allowed on YouTube.
	Harmful or dangerous content YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death.
	Suicide & self-injury We do not allow content on YouTube that promotes suicide, self-harm, or is intended to shock or disgust users.
Online piracy • Pirating, Copyright infringement, & Counterfeiting	Fake Engagement YouTube doesn't allow anything that artificially increases the number of views, likes, comments, or other metric either through the use of automatic systems or by serving up videos to unsuspecting viewers. Additionally, content that solely exists to incentivize viewers for engagement (views, likes, comments, etc) is prohibited.
	Impersonation: Content intended to impersonate a person or channel is not allowed on YouTube. YouTube also enforces trademark holder rights. When a channel, or content in the channel, causes confusion about the source of goods and services advertised, it may not be allowed.
	Sale of illegal or regulated goods or services Content intended to sell certain regulated goods and services is not allowed on YouTube. Such as: Counterfeit documents or currency
	YouTube's Terms of Service Also covered in YouTube's Terms of Service
Hate speech & acts of aggression • Behavior or content that incites hatred, promotes violence, vilifies, or dehumanizes groups or individuals based on race, ethnicity, gender, sexual orientation, gender identity, age, ability, nationality, religion, caste, victims and survivors of violent acts and their kin, immigration status, or serious disease sufferers.	Hate speech Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability (including chronic or lifelong diseases), Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust • Excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult.	Violent or Graphic Content Violent or gory content intended to shock or disgust viewers, or content encouraging others to commit violent acts are not allowed on YouTube.
	Age Restriction Sometimes content doesn't violate our policies, but it may not be appropriate for viewers under 18. In these cases, we may place an age-restriction on the video. This policy applies to videos, video descriptions, custom thumbnails, live streams, and any other YouTube product or feature. For example, this can include content with vulgur language.
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol • Promotion or sale of illegal drug use - including abuse of prescription drugs. Federal jurisdiction applies, but allowable where legal local jurisdiction can be effectively managed • Promotion and advocacy of Tobacco and e-cigarette (Vaping) & Alcohol use to minors	Sale of illegal or regulated goods or services Content intended to sell certain regulated goods and services is not allowed on YouTube. Such as: controlled narcotics and other drugs, nicotine, including vaping products, pharmaceuticals without a prescription, unlicensed medial services
	Harmful or dangerous content YouTube doesn't allow content that encourages dangerous or illegal activities that risk serious physical harm or death.
Spam or Harmful Content • Malware/Phishing	Spam deceptive practices & scams 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.
Terrorism • Promotion and advocacy of graphic terrorist activity involving defamation, physical and/or emotional harm of individuals, communities, and society	Violent criminal organizations Content intended to praise, promote, or aid violent criminal organizations is not allowed on YouTube. These organizations are not allowed to use YouTube for any purpose, including recruitment.
Debated Sensitive Social Issue • Insensitive, irresponsible and harmful treatment of debated social issues and related acts that demean a particular group or incite great conflict	
Other	COVID-19 misinfo policy YouTube doesn't allow content about COVID-19 that poses a serious risk of egregious harm. YouTube doesn't allow content that spreads medical misinformation that contradicts local health authorities' or the World Health Organization's (WHO) medical information about COVID-19. This is limited to content that contradicts WHO or local health authorities' guidance on: Treatment , Prevention, Diagnostic, Transmission, Social distancing and self isolation guidelines, The existence of COVID-19

FACEBOOK

We want Facebook and Instagram to be places where people have a voice. To create conditions where everyone feels comfortable expressing themselves, we must also protect our community's safety, privacy, dignity and authenticity. This is why we have Community Standards on Facebook and Community Guidelines on Instagram, that define what content is and is not allowed. We don't allow anything that goes against these policies, and we invest in technology, processes and people to help us act quickly 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.

The first step in mitigating harm is to fully understand how and when it occurs. We publish our Community Standards Enforcement Report on a quarterly basis to more effectively track our progress and demonstrate our continued commitment to making Facebook and Instagram safe and inclusive. We develop metrics to examine how effectively we enforce our policies, prioritize ways we can do better and hold ourselves accountable to the billions of people who use our services. Since we use these metrics for our own internal tracking, they represent our best attempt to fairly reflect how effectively we enforce our policies. The report measures:

- Prevalence How prevalent were violation views on our services?
- Content Actioned How much content did we take action on?
- Proactive Rate How much violating content did we find before users reported it?
- Appealed Content How much of the content we actioned did people appeal?
- Restored Content How much content did we restore after taking action on it, before or after an appeal?

We use prevalence to judge how we are doing at enforcement. This metric measures the percentage of times that violating content is seen on our platform, and it matters because it captures violating content that is seen, either because we missed it or because users saw it before we removed it. 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.

For more details about our processes, methodologies and how we arrived at the numbers, you can read our companion guide.

General trends for 2020

We've spent the last few years building tools, teams and technologies to help keep people safe from harmful content. So when the COVID-19 crisis emerged, we had the tools and processes in place to move quickly and we were able to continue finding and removing content that violates our policies. When we temporarily sent our content reviewers home due to the COVID-19 pandemic, we increased our reliance on these automated systems and prioritized high-severity content for our teams to review in order to continue to keep our apps safe. While our technology for identifying and removing violating content is improving, there will continue to be areas where we rely on people to both review content and train our technology. Other impacts from COVID-19 include:

- We couldn't always offer the option to appeal content decisions and account removals, as reflected through Q2-Q4 data. 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.
- We prioritized removing harmful content over measuring our efforts, so we were not able to calculate the prevalence for violent & graphic content and adult nudity & sexual activity in our Q2 data. We resumed calculating prevalence in Q3, but our prevalence estimates are specific to September of 2020 when we regained some review capacity.
- Between March and October of 2020, we removed more than 12 million pieces of content on Facebook and Instagram for containing misinformation that may lead to imminent physical harm such as content relating to fake preventative measures or exaggerated cures for COVID-19. In Q4 of 2020, that number increased to over 1 million pieces of content were removed on Facebook and Instagram globally for containing misinformation on COVID-19 that may lead to imminent physical harm.

FACEBOOK

Our ongoing commitment and investments in AI have enabled us to show improvements across both content actioned and proactive rate across many policy areas, for example between Q4 2019 and Q4 2020 we increased the volume of hate speech we took action against by 389%. We improved our proactive technologies which helped us detect and remove content that is identical or nearly identical to existing violations in our database (in areas like violent & graphic content, suicide & self injury, and hate speech). The expansion of languages including Spanish, Arabic, Portuguese and Indonesian also contributed to improvements.

Overall trends Q4 2020 (our latest report)

In our Q4 2020 report, we demonstrated improvements in prevalence on Facebook. From Q3 to Q4 2020, hate speech prevalence dropped from 0.10-0.11% to 0.07-0.08%, or 7 to 8 views of hate speech for every 10,000 views of content. The prevalence of violent and graphic content also dropped from 0.07% to 0.05%, and adult nudity content dropped from 0.05-0.06% to 0.03-0.04%. These improvements were mainly due to changes we made to reduce potentially problematic content in News Feed. Each post is ranked by processes that take into account a combination of integrity signals, such as how likely a piece of content is to violate our policies, as well as signals we receive from people, such as from surveys or actions they take on our platform like hiding or reporting posts. Improving how we use these signals helps tailor News Feed to each individual's preferences and also reduces the number of times we display posts that later may be determined to violate our policies.

On Facebook, content actioned on hate speech increased from 22.1 million pieces of content in Q3 2020 to 26.9 million in Q4 2020, primarily due to improving our proactive detection technology for the Arabic and Spanish languages. We also expanded automation for the Portuguese language, which continued to drive enforcement in Q4. Our proactive rate increased from 94.7% to 97.1% for these same reasons.

Our proactive rate improved in other problem areas, most notably bullying and harassment. Improvement on both Facebook (3.5M pieces of content to 6.3M) and Instagram (2.6M pieces of content to 5M) were driven by increasing our automation abilities and improving our technology to detect and remove more English language comments. This, in addition to regaining some manual review capacity in Q3, helped our proactive rate increase on both Facebook (from 26.4% to 48.8%), and Instagram (from 54.8% to 80%). Content actioned also increased in the areas of organized hate and restricted goods: firearms on Facebook and Instagram, as well as terrorism on Instagram, primarily driven by improvements to our proactive detection technology in Q3 and Q4.

Improvements to our AI in areas where nuance and context are essential, such as hate speech or bullying and harassment, helped us better scale our efforts to keep people safe. We use AI to help prioritize content for review, so our reviewers can focus on content that poses the most harm and spend more time training and measuring the quality of our automated systems.

We're slowly continuing to regain our content review workforce globally, though we anticipate our ability to review content will be impacted by COVID-19 until a vaccine is widely available. With limited capacity, we prioritize the most harmful content for our teams to review, such as suicide and self-injury content.

2021 Roadmap

This year, we plan to share additional metrics on Instagram and add new policy categories on Facebook. We're also working to make our enforcement data easier for people to understand by making these reports more interactive. Our goal is to lead the technology industry in transparency, and we'll continue to share more enforcement metrics as part of this effort. We also believe that no company should grade its own homework. Last year, we committed to undertaking an independent, third-party audit of our content moderation systems to validate the numbers we publish, and we'll begin this process this year.

We will continue building on this progress and improving our technology and enforcement efforts to keep harmful content off of our apps.

Question 1: How safe is the platform for consumers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	**0.05%-0.06%**	**0.03%-0.04%**	**0.05%-0.06%**	Did not report	As calculating prevalence depends on manual review, prevalence for Adult Nudity and Sexual Activity was unavailable for Q2 2020, due to a temporary reduction in our review capacity as a result of COVID-19 because we prioritized removing harmful content over measuring certain efforts.
Arms & Ammunition	Regulated Goods: Firearms	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	**0.07%**	**0.05%**	**0.07%-0.08%**	Did not report	Prevalence of Violent and Graphic Content was 0.05% of views in Q4 2020, which marks a decrease from Q3. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (As calculating prevalence depends on manual review, prevalence for violent & graphic content was unavailable for Q2 2020, due to a temporary reduction in our review capacity as a result of COVID-19 - we resumed calculating prevalence in Q3). We do not yet report prevalence of Bullying and Harassment. Our methodology to calculate prevalence for Bullying and Harassment would need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context, and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
	Child Nudity and Sexual Exploitation	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
	Suicide and Self-Injury	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
Death, Injury or Military Conflict	Violent and Graphic Content	**0.07%**	**0.05%**	**0.07%-0.08%**	Did not report	Prevalence of Violent and Graphic Content was 0.05% of views in Q4 2020, which marks a decrease from Q3. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (As calculating prevalence depends on manual review, prevalence for violent & graphic content was unavailable for Q2 2020, due to a temporary reduction in our review capacity as a result of COVID-19 - we resumed calculating prevalence in Q3).

Question 1: How safe is the platform for consumers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
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	**0.10%-0.11%**	**0.07%-0.08%**	N/A	N/A	Hate Speech: Prevalence was between 0.07% and 0.08% of views in Q4 2020, which marks a decrease from Q3 where it was between 0.10%–0.11%. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (Prevalence of Hate Speech was first introduced in our Q3 report.) We do not yet report prevalence of Bullying and Harassment. Our methodology to calculate prevalence for Bullying and Harassment would need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context, and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	**0.10%-0.11%**	**0.07%-0.08%**	N/A	N/A	Hate Speech: Prevalence was between 0.07% and 0.08% of views in Q4 2020, which marks a decrease from Q3 where it was between 0.10%–0.11%. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (Prevalence of Hate Speech was first introduced in our Q3 report.) We do not yet report prevalence of Bullying and Harassment. Our methodology to calculate prevalence for Bullying and Harassment would need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context, and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
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%**	

Question 1: How safe is the platform for consumers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
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 0.07%**	**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	**0.10%-0.11%**	**0.07%-0.08%**	N/A	N/A	Hate Speech: Prevalence was between 0.07% and 0.08% of views in Q4 2020, which marks a decrease from Q3 where it was between 0.10%–0.11%. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (Prevalence of Hate Speech was first introduced in our Q3 report.) We do not yet report prevalence of Bullying and Harassment. Our methodology to calculate prevalence for Bullying and Harassment would need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context, and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	**0.05%-0.06%**	**0.03%-0.04%**	**0.05%-0.06%**	Did not report	As calculating prevalence depends on manual review, prevalence for Adult Nudity and Sexual Activity was unavailable for Q2 2020, due to a temporary reduction in our review capacity as a result of COVID-19 because we prioritized removing harmful content over measuring certain efforts.
Arms & Ammunition	Regulated Goods: Firearms	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	**0.07%**	**0.05%**	**0.07%-0.08%**	Did not report	Prevalence of Violent and Graphic Content was 0.05% of views in Q4 2020, which marks a decrease from Q3. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (As calculating prevalence depends on manual review, prevalence for violent & graphic content was unavailable for Q2 2020, due to a temporary reduction in our review capacity as a result of COVID-19 - we resumed calculating prevalence in Q3). We do not yet report prevalence of Bullying and Harassment. Our methodology to calculate prevalence for Bullying and Harassment would need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context, and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
	Child Nudity and Sexual Exploitation	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
	Suicide and Self-Injury	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
Death, Injury or Military Conflict	Violent and Graphic Content	**0.07%**	**0.05%**	**0.07%-0.08%**	Did not report	Prevalence of Violent and Graphic Content was 0.05% of views in Q4 2020, which marks a decrease from Q3. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (As calculating prevalence depends on manual review, prevalence for violent & graphic content was unavailable for Q2 2020, due to a temporary reduction in our review capacity as a result of COVID-19 - we resumed calculating prevalence in Q3).

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
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	**0.10%-0.11%**	**0.07%-0.08%**	N/A	N/A	Hate Speech: Prevalence was between 0.07% and 0.08% of views in Q4 2020, which marks a decrease from Q3 where it was between 0.10%–0.11%. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (Prevalence of Hate Speech was first introduced in our Q3 report.) We do not yet report prevalence of Bullying and Harassment. Our methodology to calculate prevalence for Bullying and Harassment would need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context, and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	**0.10%-0.11%**	**0.07%-0.08%**	N/A	N/A	Hate Speech: Prevalence was between 0.07% and 0.08% of views in Q4 2020, which marks a decrease from Q3 where it was between 0.10%–0.11%. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (Prevalence of Hate Speech was first introduced in our Q3 report.) We do not yet report prevalence of Bullying and Harassment. Our methodology to calculate prevalence for Bullying and Harassment would need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context, and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
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%**	

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
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 0.07%**	**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	**0.10%-0.11%**	**0.07%-0.08%**	N/A	N/A	Hate Speech: Prevalence was between 0.07% and 0.08% of views in Q4 2020, which marks a decrease from Q3 where it was between 0.10%–0.11%. This was due to ranking changes to personalize content for users and reduce problematic content in News Feed. (Prevalence of Hate Speech was first introduced in our Q3 report.) We do not yet report prevalence of Bullying and Harassment. Our methodology to calculate prevalence for Bullying and Harassment would need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context, and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	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 Facebook or Instagram

		Latest Period						Previous Period					
GARM Category	Relevant Policy	Q3 2020			Q4 2020			Q1 2020			Q2 2020		
		Content Removed	% Proactive	Views & Accounts	Content Removed	% Proactive	Views & Accounts	Content Removed	% Proactive	Views & Accounts	Content Removed	% Proactive	Views & Accounts
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	36.6m	98.2%	N/A	28m	98.1%	N/A	39.5m	99.2%	N/A	35.8m	98.4%	N/A
Arms & Ammunition	Regulated Goods: Firearms	1.1m	91.5%	N/A	1.3m	92.2%	N/A	1.4m	92.9%	N/A	1.3m	91.8%	N/A
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	19.2m	99.5%	N/A	16m	99.5%	N/A	25.3m	99.0%	N/A	15m	99.5%	N/A
	Bullying and Harassment	3.5m	26.4%	N/A	6.3m	48.8%	N/A	2.3m	15.6%	N/A	2.4m	13.3%	N/A
	Child Nudity and Sexual Exploitation	12.4m	99.4%	N/A	5.4m	98.8%	N/A	8.6m	99.5%	N/A	9.5m	99.2%	N/A
	Suicide and Self-Injury	1.3m	98.0%	N/A	2.5m	92.8%	N/A	1.7m	97.6%	N/A	910k	98.0%	N/A
Death, Injury or Military Conflict	Violent and Graphic Content	19.2m	99.5%	N/A	16m	99.5%	N/A	25.3m	99.0%	N/A	15m	99.5%	N/A
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	N/A	N/A	1.1m	83.4%	N/A	1.4m	82.1%	N/A
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	N/A	N/A	348.4k	74.7%	N/A	480.3k	76.9%	N/A
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	N/A	N/A	88.4k	59.2%	N/A	123.2k	59.1%	N/A
Hate speech & acts of aggression	Hate Speech	22.1m	94.7%	N/A	26.9m	97.1%	N/A	9.5m	89.3%	N/A	22.5m	94.7%	N/A
	Bullying and Harassment	3.5m	26.4%	N/A	6.3m	48.8%	N/A	2.3m	15.6%	N/A	2.4m	13.3%	N/A
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	22.1m	94.7%	N/A	26.9m	97.1%	N/A	9.5m	89.3%	N/A	22.5m	94.7%	N/A
	Bullying and Harassment	3.5m	26.4%	N/A	6.3m	48.8%	N/A	2.3m	15.6%	N/A	2.4m	13.3%	N/A
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	4.7m	96.9%	N/A	4.3m	97.3%	N/A	7.9m	99.1%	N/A	5.8m	98.1%	N/A
Spam or Harmful Content	Spam	1.9b	99.9%	N/A	1b	99.8%	N/A	1.9b	99.9%	N/A	1.4b	99.8%	N/A
Terrorism	Dangerous Organizations: Terrorism	9.7m	99.7%	N/A	8.6m	99.8%	N/A	6.3m	99.2%	N/A	8.7m	99.6%	N/A
	Dangerous Organizations: Organized Hate	4m	97.5%	N/A	6.4m	98.3%	N/A	4.7m	96.7%	N/A	4m	96.9%	N/A
Debated Sensitive Social Issue	Hate Speech	22.1m	94.7%	N/A	26.9m	97.1%	N/A	9.5m	89.3%	N/A	22.5m	94.7%	N/A
	Bullying and Harassment	3.5m	26.4%	N/A	6.3m	48.8%	N/A	2.3m	15.6%	N/A	2.4m	13.3%	N/A

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

Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Facebook or Instagram

GARM Category	Relevant Policy	Comments
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	Adult Nudity and Sexual Activity: Content actioned decreased from 36.6 million pieces of content in Q3 2020 to 28 million in Q4 2020. In late-Q3, we adjusted our proactive detection technology for photos and videos as we continue to work on improving precision, which resulted in fewer content removals.
Arms & Ammunition	Regulated Goods: Firearms	Regulated Goods: Firearms: Content actioned increased from 1.1 million pieces of content in Q3 2020 to 1.3 million in Q4 2020. In November, we made an improvement to our proactive detection technology. As a result, the amount of content we took action on increased in Q4.
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	Violent and Graphic Content: Content actioned decreased in Q2 due to impacts from COVID-19 meaning fewer content reviewers, who are essential in our continued efforts to increase enforcement in such sensitive areas. The increase in Q3 was due to updating our use of proactive detection technology, in addition to making improvements to our technology that helped us detect and remove more potential violations.
	Bullying and Harassment	Bullying and Harassment: Content actioned increased from 3.5 million pieces of content in Q3 2020 to 6.3 million in Q4 2020, and our proactive rate increased from 26.4% to 48.8%. This was driven by increasing our automation abilities and improving our technology to detect and remove more English language comments, in addition to regaining some manual review capacity in Q3. (In Q2 we also increased our automation abilities and made improvements to our proactive detection technology for the English language.)
	Child Nudity and Sexual Exploitation	Child Nudity and Sexual Exploitation: Content actioned decreased from 12.4 million pieces of content in Q3 2020 to 5.4 million in Q4 2020, partly due to a few pieces of old, violating content that were shared widely in Q2 and Q3, which we proactively detected and removed in August. This marks a return to Q2 levels following an increase in enforcement in Q3. In mid-November, we also made changes to our internal systems that impacted enforcement for child nudity and sexual exploitation. These changes were made to our technology for detecting and removing content that is identical or near-identical to existing violations in our database. We discovered a technical issue after implementing these changes, addressed it and are working to catch content we may have missed.
	Suicide and Self-Injury	Suicide and Self Injury: Content actioned increased from 1.3 million pieces of content in Q3 2020 to 2.5 million in Q4 2020. This was driven by regaining some manual review capacity in September, which enabled us to detect and remove more photos and videos using our mediamatching technology. (Content actioned decreased in Q2, due to impacts from COVID-19 meaning fewer content reviewers, who are essential in our continued efforts to increase enforcement in such sensitive areas.)
Death, Injury or Military Conflict	Violent and Graphic Content	Violent and Graphic Content: Content actioned decreased in Q2 due to impacts from COVID-19 meaning fewer content reviewers, who are essential in our continued efforts to increase enforcement in such sensitive areas. The increase in Q3 was due to updating our use of proactive detection technology, in addition to making improvements to our technology that helped us detect and remove more potential violations.
Online piracy	Intellectual Property: Copyright	We report this metric monthly in a 6 month report. Our current report has data for January - June 2020. These numbers reflect the total amount of content that was removed based on an IP report. On Facebook, this includes everything from individual posts, photos, videos or advertisements to profiles, Pages, groups and events. Our proactive rate figure here constitutes the volume of content removed in response to an IP report relative to the volume of content reported, reflected as a percentage. In prior transparency reports, the Removal Rate constituted the percentage of total IP reports that resulted in some or all reported content being removed. Beginning in the July 2019 reporting period, we have adjusted the way we calculate Removal Rate to reflect the percentage of reported content removed, rather than the percentage of reports resulting in removals. Because a single IP report can identify multiple pieces of content, this figure offers a more complete picture of the total content removed from the platform based on an IP report.
	Intellectual Property: Counterfeit	
	Intellectual Property: Trademark	

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 or Instagram

GARM Category	Relevant Policy	Comments
Hate speech & acts of aggression	Hate Speech	Hate Speech: content actioned increased from 22.1 million pieces of content in Q3 2020 to 26.9 million in Q4 2020, and our proactive rate increased from 94.7% to 97.1%, primarily due to improving our proactive detection technology for the Arabic and Spanish languages. Our content actioned and proactive rates were also impacted by our expansion of automation for the Portuguese language, which continued to drive enforcement in Q4. (Starting in Q1, we made improvements to our proactive detection technology and expanded automation to the Spanish, Arabic and Indonesian languages. In Q2, we followed up by expanding automation to the English, Spanish and Burmese languages, which helped us detect and remove more content.)
	Bullying and Harassment	Bullying and Harassment: Content actioned increased from 3.5 million pieces of content in Q3 2020 to 6.3 million in Q4 2020, and our proactive rate increased from 26.4% to 48.8%. This was driven by increasing our automation abilities and improving our technology to detect and remove more English language comments, in addition to regaining some manual review capacity in Q3. (In Q2 we also increased our automation abilities and made improvements to our proactive detection technology for the English language.)
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 22.1 million pieces of content in Q3 2020 to 26.9 million in Q4 2020, and our proactive rate increased from 94.7% to 97.1%, primarily due to improving our proactive detection technology for the Arabic and Spanish languages. Our content actioned and proactive rates were also impacted by our expansion of automation for the Portuguese language, which continued to drive enforcement in Q4. (Starting in Q1, we made improvements to our proactive detection technology and expanded automation to the Spanish, Arabic and Indonesian languages. In Q2, we followed up by expanding automation to the English, Spanish and Burmese languages, which helped us detect and remove more content.)
	Bullying and Harassment	Bullying and Harassment: Content actioned increased from 3.5 million pieces of content in Q3 2020 to 6.3 million in Q4 2020, and our proactive rate increased from 26.4% to 48.8%. This was driven by increasing our automation abilities and improving our technology to detect and remove more English language comments, in addition to regaining some manual review capacity in Q3. (In Q2 we also increased our automation abilities and made improvements to our proactive detection technology for the English language.)
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	
Spam or Harmful Content	Spam	
Terrorism	Dangerous Organizations: Terrorism	Dangerous Organizations: Terrorism: 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 increased from 4 million pieces of content in Q3 2020 to 6.4 million in Q4 2020, primarily driven by improvements to our proactive detection technology in Q4.
Debated Sensitive Social Issue	Hate Speech	Hate Speech: content actioned increased from 22.1 million pieces of content in Q3 2020 to 26.9 million in Q4 2020, and our proactive rate increased from 94.7% to 97.1%, primarily due to improving our proactive detection technology for the Arabic and Spanish languages. Our content actioned and proactive rates were also impacted by our expansion of automation for the Portuguese language, which continued to drive enforcement in Q4. (Starting in Q1, we made improvements to our proactive detection technology and expanded automation to the Spanish, Arabic and Indonesian languages. In Q2, we followed up by expanding automation to the English, Spanish and Burmese languages, which helped us detect and remove more content.)
	Bullying and Harassment	Bullying and Harassment: Content actioned increased from 3.5 million pieces of content in Q3 2020 to 6.3 million in Q4 2020, and our proactive rate increased from 26.4% to 48.8%. This was driven by increasing our automation abilities and improving our technology to detect and remove more English language comments, in addition to regaining some manual review capacity in Q3. (In Q2 we also increased our automation abilities and made improvements to our proactive detection technology for the English language.)

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

Authorized Metric: Appeals, Reinstatements

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

		Latest Period						Previous Period						
GARM Category	Relevant Policy	Q3 2020			Q4 2020			Q1 2020			Q2 2020			Commentary
		Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	14.8k	200	18.6k	306.5k	35.1k	281.6k	2.3m	612.9k	25.2k	12.6k	100	10.8k	The number of appeals is still lower than in previous reports because we couldn't always 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.
Arms & Ammunition	Regulated Goods: Firearms	36.1k	2.1k	7.7k	44.1k	7.1k	24.9k	109k	5.1k	4.5k	41k	4.6k	9.1k	The number of appeals is still lower than in previous reports because we couldn't always 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.
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	500	90	700	3.9k	500	4.9k	481k	119k	1.5k	700	20	200	The number of appeals is still lower than in previous reports because we couldn't always 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	14.6k	3k	393.5k	442.8k	41k	131.6k	552.6k	65.8k	1.6k	600	80	43.4k	
	Child Nudity and Sexual Exploitation	300	0	1.3k	4.6k	100	3.3k	55k	3.7k	900	40	0	70	
	Suicide and Self-Injury	0	0	100	700	100	3.7k	21.5k	3.9k	200	10	50	50	

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

Authorized Metric: Appeals, Reinstatements

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

		Latest Period						Previous Period						
GARM Category	Relevant Policy	Q3 2020			Q4 2020			Q1 2020			Q2 2020			Commentary
		Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	
Death, Injury or Military Conflict	Violent and Graphic Content	500	90	700	3.9k	500	4.9k	481k	119k	1.5k	700	20	200	The number of appeals is still lower than in previous reports because we couldn't always 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
Online piracy	Intellectual Property: Copyright Intellectual Property: Counterfeit 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	We do not report content appealed and resinstated of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
Hate speech & acts of aggression	Hate Speech	41k	4.7k	232.4	984.2k	48.2k	211.2k	1.2m	59.6k	1k	30k	4.4k	145.7k	The number of appeals is still lower than in previous reports because we couldn't always 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	14.6k	3k	393.5k	442.8k	41k	131.6k	552.6k	65.8k	1.6k	600	80	43.4k	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	41k	4.7k	232.4	984.2	48.2k	211.2k	1.2m	59.6k	1k	30k	4.4k	145.7k	The number of appeals is still lower than in previous reports because we couldn't always 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	14.6k	3k	393.5k	442.8k	41k	131.6k	552.6k	65.8k	1.6k	600	80	43.4k	

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

Authorized Metric: Appeals, Reinstatements

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

		Latest Period						Previous Period						
GARM Category	Relevant Policy	Q3 2020			Q4 2020			Q1 2020			Q2 2020			Commentary
		Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	45k	6.4k	82.9k	80.2k	58.6k	316.5k	122.4k	27.8k	11.6k	28.8k	16.9k	46.7k	The number of appeals is still lower than in previous reports because we couldn't always 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	25.9k	90	74.9m	31k	700	21.4m	2.2m	535.9k	94.5m	95.3k	24.8k	43.7k	The number of appeals is still lower than in previous reports because we couldn't always 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.
Terrorism	Dangerous Organizations: Terrorism	1.7k	100	125.6k	50.3k	4.3k	51.7k	181.6k	23k	199k	300	400	533k	The number of appeals is still lower than in previous reports because we couldn't always 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	21.6k	3.8k	123.4k	137.7k	27.2k	204.2k	232.8k	53.7k	11.3k	600	1.6k	133.6k	
Debated Sensitive Social Issue	Hate Speech	41k	4.7k	232.4	984.2k	48.2k	211.2k	1.2m	59.6k	1k	30k	4.4k	145.7k	The number of appeals is still lower than in previous reports because we couldn't always 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	14.6k	3k	393.5k	442.8k	41k	131.6k	552.6k	65.8k	1.6k	600	80	43.4k	

Question 1: How safe is the platform for consumers?

Authorized Metric: Prevalence

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

		Latest Period		**Previous Period**		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	N/A	N/A	N/A	N/A	We cannot estimate this metric right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
Arms & Ammunition	Regulated Goods: Firearms	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	N/A	N/A	N/A	N/A	Violent and Graphic Content: We cannot estimate this metric right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data. Bullying & Harassment: We cannot estimate this metric right now. Our methodology to calculate prevalence for Bullying and Harassment will need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
	Child Nudity and Sexual Exploitation	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
	Suicide and Self-Injury	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
Death, Injury or Military Conflict	Violent and Graphic Content	N/A	N/A	N/A	N/A	Violent and Graphic Content: We cannot estimate this metric 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?

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 Q3 2020	Q4 2020	Previous Period Q1 2020	Q2 2020	Commentary
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	N/A	N/A	N/A	N/A	Violent and Graphic Content: We cannot estimate this metric right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data. Bullying & Harassment: We cannot estimate this metric right now. Our methodology to calculate prevalence for Bullying and Harassment will need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	N/A	N/A	N/A	N/A	Violent and Graphic Content: We cannot estimate this metric right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data. Bullying & Harassment: We cannot estimate this metric right now. Our methodology to calculate prevalence for Bullying and Harassment will need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
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%**	

Question 1: How safe is the platform for consumers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
Spam or Harmful Content	Spam	N/A	N/A	N/A	N/A	
Terrorism	Dangerous Organizations: Terrorism	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**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	N/A	N/A	N/A	N/A	Violent and Graphic Content: We cannot estimate this metric right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data. Bullying & Harassment: We cannot estimate this metric right now. Our methodology to calculate prevalence for Bullying and Harassment will need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	N/A	N/A	N/A	N/A	We cannot estimate this metric right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
Arms & Ammunition	Regulated Goods: Firearms	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	N/A	N/A	N/A	N/A	Violent and Graphic Content: We cannot estimate this metric right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data. Bullying & Harassment: We cannot estimate this metric right now. Our methodology to calculate prevalence for Bullying and Harassment will need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment
	Bullying and Harassment	N/A	N/A	N/A	N/A	
	Child Nudity and Sexual Exploitation	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
	Suicide and Self-Injury	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	**Less than 0.05%**	
Death, Injury or Military Conflict	Violent and Graphic Content	N/A	N/A	N/A	N/A	Violent and Graphic Content: We cannot estimate this metric right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
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	N/A	N/A	N/A	N/A	Hate Speech: We cannot estimate this metric right now. Our prevalence measurement is slowly expanding to cover more languages and regions to account for cultural context and nuances for individual languages. We are still developing a global metric, although our detection and enforcement of hate speech is very broad across the world. Bullying & Harassment: We cannot estimate this metric right now. Our methodology to calculate prevalence for Bullying and Harassment will need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	N/A	N/A	N/A	N/A	Hate Speech: We cannot estimate this metric right now. Our prevalence measurement is slowly expanding to cover more languages and regions to account for cultural context and nuances for individual languages. We are still developing a global metric, although our detection and enforcement of hate speech is very broad across the world. Bullying & Harassment: We cannot estimate this metric right now. Our methodology to calculate prevalence for Bullying and Harassment will need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	N/A	N/A	
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%**	

Question 2: How safe is the platform for advertisers?

Authorized Metric: Prevalence

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

		Latest Period		Previous Period		
GARM Category	**Relevant Policy**	**Q3 2020**	**Q4 2020**	**Q1 2020**	**Q2 2020**	**Commentary**
Spam or Harmful Content	Spam	N/A	N/A	N/A	N/A	
Terrorism	Dangerous Organizations: Terrorism	**Less than 0.05%**	**Less than 0.07%**	**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	N/A	N/A	N/A	N/A	Hate Speech: We cannot estimate this metric right now. Our prevalence measurement is slowly expanding to cover more languages and regions to account for cultural context and nuances for individual languages. We are still developing a global metric, although our detection and enforcement of hate speech is very broad across the world. Bullying & Harassment: We cannot estimate this metric right now. Our methodology to calculate prevalence for Bullying and Harassment will need to be different from the way we measure the prevalence of other violations. This is because communication can be highly dependent on language and context and often reflects the nature of personal relationships. In many instances, we need a person to report this behavior to us before we can identify it as bullying or harassment.
	Bullying and Harassment	N/A	N/A	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 Facebook or Instagram

		Latest Period						Previous Period					
GARM Category	Relevant Policy	Q3 2020			Q4 2020			Q1 2020			Q2 2020		
		Content Removed	% Proactive	Views & Accounts	Content Removed	% Proactive	Views & Accounts	Content Removed	% Proactive	Views & Accounts	Content Removed	% Proactive	Views & Accounts
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	13.1m	93.5%	N/A	11.5m	97%	N/A	8.1m	93.8%	N/A	12.4m	96.1%	N/A
Arms & Ammunition	Regulated Goods: Firearms	36.5k	87.9%	N/A	70.2k	90%	N/A	54.3k	90.8%	N/A	50.7k	92.2%	N/A
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	4.1m	97.5%	N/A	5.6m	98%	N/A	2.8m	95.4%	N/A	3.1m	97.0%	N/A
	Bullying and Harassment	2.6m	54.7%	N/A	5m	80%	N/A	1.5m	35.2%	N/A	2.3m	37.7%	N/A
	Child Nudity and Sexual Exploitation	1m	96.7%	N/A	800k	98%	N/A	1m	97.7%	N/A	480k	95.9%	N/A
	Suicide and Self-Injury	1.3m	95.4%	N/A	3.4m	95%	N/A	1.3m	89.7%	N/A	303.6k	93.8%	N/A
Death, Injury or Military Conflict	Violent and Graphic Content	4.1m	97.5%	N/A	5.6m	98%	N/A	2.8m	95.4%	N/A	3.1m	97.0%	N/A
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	N/A	N/A	547.4k	89.6%	N/A	656k	85.4%	N/A
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	N/A	N/A	238.4k	90.5%	N/A	241.6k	91.9%	N/A
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	N/A	N/A	85.8k	58.2%	N/A	106.9k	60.7%	N/A
Hate speech & acts of aggression	Hate Speech	6.5m	94.8%	N/A	6.6m	95%	N/A	578k	42.9%	N/A	3.2m	84.9%	N/A
	Bullying and Harassment	2.6m	54.7%	N/A	5m	80%	N/A	1.5m	35.2%	N/A	2.3m	37.7%	N/A
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	6.5m	94.8%	N/A	6.6m	95%	N/A	578k	42.9%	N/A	3.2m	84.9%	N/A
	Bullying and Harassment	2.6m	54.7%	N/A	5m	80%	N/A	1.5m	35.2%	N/A	2.3m	37.7%	N/A
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	1.2m	93.9%	N/A	1.4m	96%	N/A	1.3m	94.7%	N/A	1.4m	94.6%	N/A
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	183.2k	99.2%	N/A	345.1k	98%	N/A	440.7k	84.8%	N/A	388.8k	98.5%	N/A
	Dangerous Organizations: Organized Hate	224.7k	77.3%	N/A	308k	69%	N/A	175.3k	68.9%	N/A	266k	74.3%	N/A
Debated Sensitive Social Issue	Hate Speech	6.5m	94.8%	N/A	6.6m	95%	N/A	578k	42.9%	N/A	3.2m	84.9%	N/A
	Bullying and Harassment	2.6m	54.7%	N/A	5m	80%	N/A	1.5m	35.2%	N/A	2.3m	37.7%	N/A

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

Authorized Metric: Content Actioned, Actors Actioned, Proactive Rate

Violating content acted upon and removed by Facebook or Instagram

GARM Category	Relevant Policy	Comments
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	Adult Nudity and Sexual Activity: content actioned and proactive rate increased in Q2, primarily driven by improvements to our proactive detection technology to detect and remove more content automatically.
Arms & Ammunition	Regulated Goods: Firearms	Regulated Goods: Firearms: Content actioned increased from 36.5K pieces of content in Q3 2020 to 70.2K in Q4 2020, due to improvements to our proactive detection technology in October. Our proactive rate increased from 87.9% to 90.3%, for the same reasons.
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	Violent and Graphic Content: Content actioned increased from 4.1 million pieces of content in Q3 2020 to 5.6 million in Q4 2020. This was driven by a piece of violating content that was shared widely in October, which we proactively detected and removed. (Our increases in Q2 were due to improvements and expansions in our proactive detection technology, which helped us detect and remove content that is identical or nearly identical to existing violations in our database.)
	Bullying and Harassment	Bullying and Harassment: Content actioned increased from 2.6M pieces of content in Q3 2020 to 5M in Q4 2020, our proactive rate increased from 54.7% to 80%. This was driven by improving our technology to detect and remove more English language comments, in addition to regaining some manual review capacity in Q3. (Our increase in Q2 was driven by expanding our proactive detection technologies for the English and Spanish languages.)
	Child Nudity and Sexual Exploitation	Child Nudity and Sexual Exploitation: Content actioned increased in Q3 due to expanding our proactive detection technology to detect and remove more content, in addition to regaining some manual review capacity.
	Suicide and Self-Injury	Suicide and Self-Injury: Content actioned increased from 1.3 million pieces of content in Q3 2020 to 3.4 million in Q4 2020. This was primarily driven by violating content that was shared widely and very quickly across stories and profile pictures, which we proactively detected and removed. (The increase in Q3 was primarily driven by violating content that was shared widely and quickly in July and August. We also expanded our use of technology to detect and remove content that is identical or nearly identical to existing violations in our database.)
Death, Injury or Military Conflict	Violent and Graphic Content	Violent and Graphic Content: Content actioned increased from 4.1 million pieces of content in Q3 2020 to 5.6 million in Q4 2020. This was driven by a piece of violating content that was shared widely in October, which we proactively detected and removed. (Our increases in Q2 were due to improvements and expansions in our proactive detection technology, which helped us detect and remove content that is identical or nearly identical to existing violations in our database.)
Online piracy	Intellectual Property: Copyright	We report this metric monthly in a 6 month report. Our current report has data for January - June 2020. These numbers reflect the total amount of content that was removed based on an IP report. On Instagram this could include photos, videos, advertisements or accounts. Our proactive rate figure here constitutes the volume of content removed in response to an IP report relative to the volume of content reported, reflected as a percentage. In prior transparency reports, the Removal Rate constituted the percentage of total IP reports that resulted in some or all reported content being removed. Beginning in the July 2019 reporting period, we have adjusted the way we calculate Removal Rate to reflect the percentage of reported content removed, rather than the percentage of reports resulting in removals. Because a single IP report can identify multiple pieces of content, this figure offers a more complete picture of the total content removed from the platform based on an IP report.
	Intellectual Property: Counterfeit	
	Intellectual Property: Trademark	

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 or Instagram

GARM Category	Relevant Policy	Comments
Hate speech & acts of aggression	Hate Speech	Hate Speech: Content actioned increased in Q3 with the proactive rate increasing from about 85% to 95%. This was driven in part by improving our proactive detection technology for English, Arabic and Spanish languages, and expanded automation for violating media and comments. (Our increase in Q3 partly due to expanding automation to the Arabic and Indonesian languages toward the end of Q2.)
	Bullying and Harassment	Bullying and Harassment: Content actioned increased from 2.6M pieces of content in Q3 2020 to 5M in Q4 2020, our proactive rate increased from 54.7% to 80%. This was driven by improving our technology to detect and remove more English language comments, in addition to regaining some manual review capacity in Q3. (Our increase in Q2 was driven by expanding our proactive detection technologies for the English and Spanish 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 in Q3 with the proactive rate increasing from about 85% to 95%. This was driven in part by improving our proactive detection technology for English, Arabic and Spanish languages, and expanded automation for violating media and comments. (Our increase in Q3 partly due to expanding automation to the Arabic and Indonesian languages toward the end of Q2.)
	Bullying and Harassment	Bullying and Harassment: Content actioned increased from 2.6M pieces of content in Q3 2020 to 5M in Q4 2020, our proactive rate increased from 54.7% to 80%. This was driven by improving our technology to detect and remove more English language comments, in addition to regaining some manual review capacity in Q3. (Our increase in Q2 was driven by expanding our proactive detection technologies for the English and Spanish languages.) .
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	Regulated Goods: Drugs: Content actioned increased from 1.2M pieces of content in Q3 2020 to 1.4M in Q4 2020. In November, we made an improvement to our proactive detection technology for different drug violation types, which increased the amount of content we took action on in December. Our proactive rate increased from 93.9% to 96.1%. In November, we made an improvement to our proactive detection technology for different drug violation types, which increased the amount of content we took action on in December.
Spam or Harmful Content	Spam	N/A
Terrorism	Dangerous Organizations: Terrorism	Dangerous Organizations: Terrorism: Content actioned increased from 183.2K pieces of content in Q3 2020 to 345.1K in Q4 2020. This was driven by an improvement to our proactive detection technology in November, which increased our automation abilities throughout Q4. (After world events in Q1, content actioned for terrorism decreased in Q2.)
	Dangerous Organizations: Organized Hate	Dangerous Organizations: Organized hate: Content actioned increased from 224.7K pieces of content in Q3 2020 to 308K in Q4 2020. This was driven by an improvement to our proactive detection technology in November, which increased our automation abilities throughout Q4.
Debated Sensitive Social Issue	Hate Speech	Hate Speech: Content actioned increased in Q3 with the proactive rate increasing from about 85% to 95%. This was driven in part by improving our proactive detection technology for English, Arabic and Spanish languages, and expanded automation for violating media and comments. (Our increase in Q3 partly due to expanding automation to the Arabic and Indonesian languages toward the end of Q2.)
	Bullying and Harassment	Bullying and Harassment: Content actioned increased from 2.6M pieces of content in Q3 2020 to 5M in Q4 2020, our proactive rate increased from 54.7% to 80%. This was driven by improving our technology to detect and remove more English language comments, in addition to regaining some manual review capacity in Q3. (Our increase in Q2 was driven by expanding our proactive detection technologies for the English and Spanish languages.) .

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

Authorized Metric: Appeals, Reinstatements

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

		Latest Period						Previous Period						
GARM Category	Relevant Policy	Q3 2020			Q4 2020			Q1 2020			Q2 2020			Commentary
		Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	
Adult & Explicit Sexual Content	Adult Nudity and Sexual Activity	0	30	3.3k	0	20	141k	509.1k	98.2k	10.1k	0	70	3.2k	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.
Arms & Ammunition	Regulated Goods: Firearms	0	0	200	0	0	800	4.3k	400	40	0	600	200	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.
Crime & Harmful acts to individuals and Society, Human Right Violations	Violent and Graphic Content	0	0	200	0	0	3.2k	3k	500	300	0	0	200	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	183.1k	0	0	43k	87.6k	20.9k	2.7k	0	10	12.7k	
	Child Nudity and Sexual Exploitation	0	10	700	0	0	2.9k	53.4k	16.2k	200	0	30	0	
	Suicide and Self-Injury	0	10	48k	0	0	276.5k	67.2k	4.5k	1.3k	0	0	100	

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

Authorized Metric: Appeals, Reinstatements

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

		Latest Period						Previous Period						
GARM Category	Relevant Policy	Q3 2020			Q4 2020			Q1 2020			Q2 2020			Commentary
		Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	
Death, Injury or Military Conflict	Violent and Graphic Content	0	0	200	0	0	3.2k	3k	500	300	0	0	200	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.
Online piracy	Intellectual Property: Copyright	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	We do not report content appealed and resinstated of Intellectual Property Copyright, Counterfeit, Trademark. We focus on reports submitted, action rate and content removed.
	Intellectual Property: Counterfeit	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Intellectual Property: Trademark	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
Hate speech & acts of aggression	Hate Speech	0	0	32.9k	0	0	35.7k	42.9k	8.6k	1.6k	0	20	22.7k	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.
	Bullying and Harassment	0	0	183.1k	0	0	43k	87.6k	20.9k	2.7k	0	100	12.7k	
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech	0	0	32.9k	0	0	35.7k	42.9k	8.6k	1.6k	0	20	22.7k	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.
	Bullying and Harassment	0	0	183.1k	0	0	43k	87.6k	20.9k	2.7k	0	100	12.7k	

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

Authorized Metric: Appeals, Reinstatements

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

		Latest Period						Previous Period						
GARM Category	**Relevant Policy**	**Q3 2020**			**Q4 2020**			**Q1 2020**			**Q2 2020**			**Commentary**
		Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	Appealed	reinstated	Reinstated without appeal	
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	Regulated Goods: Drugs	0	10	35.7k	0	0	62.1k	95.1k	29.8k	2.4k	0	14k	22.6k	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	N/A
Terrorism	Dangerous Organizations: Terrorism	0	0	100	0	0	10	8.4k	500	200	0	0	90	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	1k	0	0	1.8k	3.9k	500	300	0	0	900	
Debated Sensitive Social Issue	Hate Speech	0	0	32.9k	0	0	35.7k	42.9k	8.6k	1.6k	0	20	22.7k	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.
	Bullying and Harassment	0	0	183.1k	0	0	43k	87.6k	20.9k	2.7k	0	100	12.7k	

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	
Arms and Ammunition	Violence and Criminal Behavior (includes policies on: Violence and Incitement, Dangerous Individuals	and Organizations, Coordinating Harm and Publicizing Crime, Regulated Goods and Fraud and Deception)
Crime and Harmful Acts to Individuals and Society and Human Right Violations	Violence and Criminal Behavior Bullying and Harassment Violent and Graphic Content	Child Sexual Exploitation, Abuse and Nudity Suicide and Self-Injury Cruel and Insensitive
Death, Injury or Military Conflict	Violence and Criminal Behavior Violent and Graphic Content	Cruel and Insensitive
Online Piracy	Intellectual Property Fraud and Deception	
Hate Speech and Acts of Aggression	Hate Speech Bullying and Harrassment	Dangerous Individuals and Organizations Cruel and Insensitive
Obscenity and Profanity, including language, gestures and explicitly gory, graphic or repulsive content intended to shock and disgust	Hate Speech Bullying and Harrassment Cruel and Insensitive	
Illegal Drugs/Tobacco/ E-cigarettes/Vaping/Alcohol	Regulated Goods	
Spam or Harmful Content	Cybersecurity Spam	
Terrorism	Dangerous Individuals and Organizations	
Debated Sensitive Social Issues	Hate Speech Bullying and Harrassment	Cruel and Insensitive

Additional policies not covered

Floor focuses online and not on offline/real-world fraud / Fraud and Deception

Floor covers explicit injury, but promoting self-injury and eating disorders is not covered / Suicide and Self-Injury

Floor does not include census and voter interference/fraud / Coordinating Harm and Publicizing Crime

Floor does not include coverage for exploitation / Sexual Exploitation of Adults

Policies Floor does not address

Privacy Violations and Image Privacy Rights
Misrepresentation
Inauthentic Behavior
False News
Manipulated Media
Memorialization
User Requests
Additional Protections for Minors

Facebook Policy

Fraud and Deception
Suicide and Self-Injury
Coordinating Harm and Publicizing Crime
Sexual Exploitation of Adults

Twitter

Twitter's fundamental belief that the open exchange of information can have a positive global impact inspired us to launch one of the industry's first transparency reports back in 2012. This transparency is a key tenet of our efforts to advance and build trust for the Open Internet.

A lot has changed since 2012. It is now more important than ever that we also shine a light on our own practices, including enforcement of the Twitter Rules and our ongoing work to disrupt global state-backed information operations. The public, policymakers, and the advertising community want to be better informed about our actions and we recognize these calls for greater transparency.

In August 2020, we evolved our biannual reports into a more comprehensive Twitter Transparency Center covering a broader array of our transparency efforts. The center includes sections covering information requests, removal requests, copyright notices, trademark notices, email security, Twitter Rules enforcement, platform manipulation, and state-backed information operations.

The metrics in this report from GARM reflect the enforcement of the Twitter Rules, which apply to everyone who uses Twitter. Our rules exist to ensure all 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 monetization on Twitter in due course, and we're working with third parties to build independent brand safety reporting solutions that will provide additional insights aligned with the GARM framework.

Our latest Twitter Transparency Report includes data from January 1, 2020, through June 30, 2020. During this reporting period, the COVID-19 pandemic severely impacted business operations around the world, disrupting our content moderation work and the way in which teams assess content and enforce our policies. In response to these disruptions, we increased our use of machine learning and automation to take a wide range of actions on potentially abusive and misleading content, whilst continually focusing human review in areas where the likelihood of harm was the greatest.

In March 2020, we launched a COVID-19 misleading information policy to further protect the health of the public conversation. During this reporting period, our teams took enforcement action against 4,658 accounts for violations of this policy. As we've further invested in technology, our automated systems challenged 4.5 million accounts that were targeting discussions around COVID-19 with spammy or manipulative behaviors. We've since expanded this policy to address misleading information about COVID-19 vaccines, which is not reflected in the H1 2020 data in this report.

Twitter discloses state-backed actors' attempts to disrupt the conversation on the service. During this reporting period, we took action on more than 52,000 accounts that we reliably attributed to information operations originating from countries around the world.

There will always be more work to do in this space, and we'll continue to provide biannual Twitter Transparency Reports that offer more clarity into our operations and work to protect the public conversation. We also recognize the importance of measuring prevalence of certain content on Twitter, and we have begun a multi-year initiative to enable us to provide more consistent transparency on these issues. We are committed to providing meaningful transparency to the public through ongoing improvements and updates to our transparency center.

Question 1: How safe is the platform for consumers?

Next best measure: Content Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Policy	Latest Period			Previous Period			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Adult & explicit sexual content	Non-consensual nudity	9,227	2,217	15,546	17,769	4,960	36,664	Twitter recognizes the importance of measuring the prevalence of certain content, and we have begun a multi-year initiative to enable us to provide more consistent transparency on these issues. We look forward to sharing more details in due course.
	Sensitive media	167,697	10,706	170,670	146,356	17,525	150,767	
	Child sexual exploitation	444,781	438,809	10,343	264,625	257,768	11,026	
Arms & ammunition	Illegal or certain regulated goods or services	56,513	54,070	16,663	60,807	56,236	37,361	
Crime & harmful acts to individuals and society, human right violations	Violence	23,835	17,493	24,161	45,447	34,196	49,172	
	Abuse/harassment	398,057	72,139	609,253	602,622	94,608	944,209	
Death, injury or military conflict	Promoting suicide or self-harm	64,610	3,302	73,656	127,736	5,612	142,930	
Online piracy	Copyright	153,325	1,216,626	1,120,593	122,758	641,715	133,920	
	Trademark	Trademark Notices: 14,917			Trademark Notices: 14,369			
Hate speech & acts of aggression	Hateful conduct	635,415	127,954	955,212	970,109	170,994	1,445,469	
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	167,697	10,706	170,670	146,356	17,525	150,767	
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services	56,513	54,070	16,663	60,807	56,236	37,361	
Spam or harmful content	Private information	25,756	3,390	37,614	22,523	3,527	34,564	
	Impersonation	131,136	120,066	12,484	181,800	168,061	16,505	
	Platform manipulation	Anti-Spam Challenges Issued: 88,008,270			Anti-Spam Challenges Issued: 135,676,973			
Terrorism	Terrorism/violent extremism	90,684	90,684		86,799	86,799		
Debated sensitive social issues	N/A							

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 Policy	Latest Period			Previous Period			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Adult & explicit sexual content	Non-consensual nudity	9,227	2,217	15,546	17,769	4,960	36,664	In December 2020, Twitter announced that we've selected DoubleVerify (DV) and Integral Ad Science (IAS) to be Twitter's preferred partners for providing independent reporting on the context in which ads appear on Twitter. We see this as an opportunity to build solutions that will give advertisers a better understanding of the types of content that appear adjacent to their ads, helping them make informed decisions to reach their marketing goals. We also hope to provide additional insights for marketers on campaign delivery specific to the GARM framework through these partnerships. We look forward to partnering with both DV and IAS to create custom solutions for our unique platform and we intend to start testing solutions in early 2021.
	Sensitive media	167,697	10,706	170,670	146,356	17,525	150,767	
	Child sexual exploitation	444,781	438,809	10,343	264,625	257,768	11,026	
Arms & ammunition	Illegal or certain regulated goods or services	56,513	54,070	16,663	60,807	56,236	37,361	
Crime & harmful acts to individuals and society, human right violations	Violence	23,835	17,493	24,161	45,447	34,196	49,172	
	Abuse/harassment	398,057	72,139	609,253	602,622	94,608	944,209	
Death, injury or military conflict	Promoting suicide or self-harm	64,610	3,302	73,656	127,736	5,612	142,930	
Online piracy	Copyright	153,325	1,216,626	1,120,593	122,758	641,715	133,920	
	Trademark	Trademark Notices: 14,917			Trademark Notices: 14,369			
Hate speech & acts of aggression	Hateful conduct	635,415	127,954	955,212	970,109	170,994	1,445,469	
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	167,697	10,706	170,670	146,356	17,525	150,767	
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services	56,513	54,070	16,663	60,807	56,236	37,361	
Spam or harmful content	Private information	25,756	3,390	37,614	22,523	3,527	34,564	
	Impersonation	131,136	120,066	12,484	181,800	168,061	16,505	
	Platform manipulation	Anti-Spam Challenges Issued: 88,008,270			Anti-Spam Challenges Issued: 135,676,973			
Terrorism	Terrorism/violent extremism	90,684	90,684		86,799	86,799		
Debated sensitive social issues	N/A							

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

Authorized Metric: Content Removals and Account Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Policy	Latest Period			Previous Period			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Adult & explicit sexual content	Non-consensual nudity	9,227	2,217	15,546	17,769	4,960	36,664	There was a 48% decrease in the number of accounts actioned for violations of our non-consensual nudity policy during the latest reporting period.
	Sensitive media	167,697	10,706	170,670	146,356	17,525	150,767	There was a 15% increase in the number of accounts actioned for violations of our sensitive media policy during the latest reporting period.
	Child sexual exploitation	444,781	438,809	10,343	264,625	257,768	11,026	We do not tolerate child sexual exploitation on Twitter. When we are made aware of child sexual exploitation media, including links to images of or content promoting child exploitation, the material will be removed from the site without further notice and reported to The National Center for Missing & Exploited Children ("NCMEC"). People can report content that appears to violate the Twitter Rules regarding Child Sexual Exploitation via our web form or through in-app reporting. 438,809 unique accounts were suspended during this reporting period for violating Twitter policies prohibiting child sexual exploitation. 91% of those accounts were proactively identified by employing internal proprietary tools and industry hash sharing initiatives. These tools and initiatives support our efforts to surface potentially violative content for further review and, if appropriate, removal.
Arms & ammunition	Illegal or certain regulated goods or services	56,513	54,070	16,663	60,807	56,236	37,361	There was a 7% decrease in the number of accounts actioned for violations of our illegal or certain regulated goods or services policy during the latest reporting period.
Crime & harmful acts to individuals and society, human right violations	Violence	23,835	17,493	24,161	45,447	34,196	49,172	There was a 48% decrease in the number of accounts actioned for violations of our violence policies during the latest reporting period.
	Abuse/harassment	398,057	72,139	609,253	602,622	94,608	944,209	There was a 34% decrease in the number of accounts actioned for violations of our abuse policy during the latest reporting period.

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

Authorized Metric: Content Removals and Account Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Policy	Latest Period			Previous Period			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Death, injury or military conflict	Promoting suicide or self-harm	64,610	3,302	73,656	127,736	5,612	142,930	There was a 49% decrease in the number of accounts actioned for violations of our suicide or self-harm policy during the latest reporting period.
Online piracy	Copyright	153,325	1,216,626	1,120,593	122,758	641,715	133,920	We report on DMCA takedown notices submitted through our web form or otherwise sent to Twitter, such as via fax or mail. We received 25% more DMCA takedown notices affecting 90% more accounts during this reporting period. There was a substantial increase in the number of Tweets (vs Tweet media) withheld in this period due to limitations in human review capabilities caused by the COVID-19 pandemic. 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.
	Trademark	Trademark Notices: 14,917			Trademark Notices: 14,369			Twitter received 4% less trademark notices, affecting 30% less accounts since our last report. We carefully review each report received under our trademark policy, and follow up with the reporter as appropriate, such as in cases of apparent fair use. We may take action on reported content if it is using another's trademark in a manner that may mislead others about its business affiliation.
Hate speech & acts of aggression	Hateful conduct	635,415	127,954	955,212	970,109	170,994	1,445,469	There was a 35% decrease in the number of accounts actioned for violations of our hateful conduct policy during the latest reporting period. Hateful conduct expanded to include a new dehumanization policy on March 5, 2020.
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Sensitive media	167,697	10,706	170,670	146,356	17,525	150,767	There was a 15% increase in the number of accounts actioned for violations of our sensitive media policy during the latest reporting period.

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

Authorized Metric: Content Removals and Account Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Policy	Latest Period			Previous Period			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol	Illegal or certain regulated goods or services	56,513	54,070	16,663	60,807	56,236	37,361	There was a 7% decrease in the number of accounts actioned for violations of our illegal or certain regulated goods or services policy during the latest reporting period.
Spam or harmful content	Private information	25,756	3,390	37,614	22,523	3,527	34,564	This reporting period saw the largest increase in the number of accounts actioned under this policy. Internal tooling improvements allowed us to increase enforcement of this policy.
	Impersonation	131,136	120,066	12,484	181,800	168,061	16,505	There was a 28% decrease in the number of accounts actioned for violations of our impersonation policy during the latest reporting period.
	Platform manipulation	Anti-Spam Challenges Issued: 88,008,270			Anti-Spam Challenges Issued: 135,676,973			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. Anti-spam challenges issued to suspected spam accounts increased substantially, by just over 54%, compared to the previous reporting period. Actions taken to counter spam tend to fluctuate for a variety of reasons, such as the volume of attempted Twitter signups, as well as the volume of spam campaigns targeting our service at any point in time.

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

Authorized Metric: Content Removals and Account Removals

Violating content acted upon and removed by Twitter

GARM Category	Relevant Policy	Latest Period			Previous Period			Commentary
		Accounts Actioned:	Accounts Suspended:	Content Removed:	Accounts Actioned:	Accounts Suspended:	Content Removed:	
Terrorism	Terrorism/violent extremism	90,684	90,684		86,799	86,799		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. Anti-spam challenges issued to suspected spam accounts increased substantially, by just over 54%, compared to the previous reporting period. Actions taken to counter spam tend to fluctuate for a variety of reasons, such as the volume of attempted Twitter signups, as well as the volume of spam campaigns targeting our service at any point in time.
Debated sensitive social issues	N/A							
Other	Civic integrity	2,351		2,710	1,721		2,146	There was a 37% increase in the number of accounts actioned for violations of our civic integrity policy during the latest reporting period. This reporting period saw an increase in the number of accounts actioned under this policy. Enforcements increased in the lead up to the US elections in November 2020.
	COVID-19 misleading information	4,658	1,751	4,647	N/A	N/A	N/A	We suspended or required the removal of content from 4,658 accounts for violations of our COVID-19 misleading information policy during the latest report period. This number does not include accounts where we applied a label or warning message,

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

Authorized Metric: Proactive Action Rate

Violating content acted upon and removed by Twitter

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & explicit sexual content				We proactively detect and action approximately 50% of content that violates our Rules today. We occasionally publish proactive detection rates for select policies at our discretion, but this is not a formal metric that we track in our Transparency Center today.
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 3: How Effective is the Platform in Enforcing Safety Policy?

Authorized Metric: Content Removals by Views

Violating content acted upon and removed by Twitter

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & explicit sexual content				Twitter does not report impressions data at this time.
Arms & ammunition				
Crime & harmful acts to individuals and society, human right violations				
Death, injury or military conflict				
Online piracy				
Hate speech & acts of aggression				
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust				
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol				
Spam or harmful content				
Terrorism				
Debated sensitive social issue				

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

Not submitted

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & explicit sexual content				Twitter does not report appeals data at this time.
Arms & ammunition				
Crime & harmful acts to individuals and society, human right violations				
Death, injury or military conflict				
Online piracy				
Hate speech & acts of aggression				
Obscenity and profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust				
Illegal drugs/tobacco/e-cigarettes/vaping/alcohol				
Spam or harmful content				
Terrorism				
Debated sensitive social issue				

TikTok

TikTok is a diverse, global community fueled by creative expression. We work to maintain an environment where everyone feels safe and welcome to create videos, find community, and be entertained. We believe that feeling safe is essential to feeling comfortable expressing yourself authentically, which is why we strive to uphold our Community Guidelines by removing accounts and content that violate them. Our goal is for TikTok to remain a place for inspiration, creativity, and joy. We are committed to being transparent about how our policies are enforced, because it helps build trust with our community and holds us accountable.

Our Transparency Reports provides visibility into the volume and nature of content removed for violating our Community Guidelines or Terms of Service. Our most recent report covers the second half of 2020 (July 1 - December 31) and includes additional information on our work to counter COVID-19 misinformation, maintain the integrity of our platform throughout global elections, and promote community well-being.

July to December 2020

In the second half of 2020, we continued our work to support our community with authoritative information about elections, COVID-19, and vaccines while we also removed misinformation related to voting, elections, public health, and more, such as elections and anti-vaccine misinformation.

Maintaining platform integrity through the US 2020 elections:

Our teams worked to safeguard the integrity of elections globally. To further these efforts, we expanded our global fact-checking partnerships, worked closely with Electoral Commissions in multiple regions, developed product features to provide our users with authoritative electoral information, and improved our internal rapid response capabilities and processes.

- Though politics and news make up a smaller amount of overall content on TikTok, and we don't accept paid political ads, we work to keep TikTok free of election misinformation and provide our community with access to authoritative information.

Countering COVID-19 and vaccine misinformation:

We make authoritative public health information available directly in our app – from our Discover page, on relevant search results, hashtags, and videos, and at our Safety Center. In our COVID-19 information hub, our community can find answers to common questions about the coronavirus and vaccines from the World Health Organization (WHO) and the Centers for Disease Control (CDC) as well as tips on staying safe.

Misinformation:

At TikTok, we work diligently to protect the integrity of our platform and take multiple approaches to help authentic content thrive. This includes prohibiting activities or content that may undermine platform integrity, such as misinformation related to civic processes or public health. Misinformation is defined as content that is inaccurate or false.

- We added fact-checking partners to additional markets and now have support in 16 languages. We've also made improvements in our ability to detect and remove fake engagement and spam.

Adult Nudity & Sexual Activities and Minor Safety:

Adult Nudity & Sexual Activities and Minor Safety remain the two most common reasons for content removal from the platform - with a decreasing volume for adult nudity and increasing volume for minor safety. Our minor safety policy is focused holistically on ways to keep minors safe from harmful or risky behavior and activities, including the possession or consumption of substances prohibited for minors, the misuse of legal substances, engagement in illegal activities, participation in activities, physical challenges, or dares that may threaten the well-being of minors.

- We've expanded our harmful activities by minors policy to further remove content that depicts minors in possession of alcohol and tobacco products (both ingestion and possession are treated equally and will be removed) as well as other behavior that could put the well-being of minors at risk.

TikTok

Harassment & Bullying & Hateful Behavior:

With highly political and emotionally charged events occurring throughout 2020, we saw an increase in hateful content and harassment & bullying on the platform. While it is still a low amount, we are especially mindful of this quickly evolving content and are working to strengthen our mechanisms to counter it, such as increasing training with moderators and interventions that promote kindness on TikTok

- The increase reflects adjustments to policies around sexual harassment, threats of hacking, and targets of bullying statements, which are now more comprehensive. Additionally we saw modest improvements in our abilities to detect harassment or bullying proactively which still remains a challenge with linguistic and cultural nuances.

Hateful Behavior:

We changed this policy from "hate speech" to its current name "hateful behavior" to take a more comprehensive approach to combatting hateful ideologies and off-platform activities.

We're proud of the progress we've made to increase visibility into our content and moderation practices. Our work is never done, and that is why we are continuously evolving our policies, building relationships with a range of experts, and investing in our technology and teams so that TikTok remains a safe place for all to express their creativity.

Question 1: How safe is the platform for consumers?

Next best measure: Policy Violation Score

Percentage volume of videos removed by policy violation

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	0.36	0.2230	Minor Safety includes anything that may perpetuate the abuse, harm, endangerment, or exploitation of minors
	Adult nudity and sexual activities	0.205	0.3090	
Arms & Ammunition	Illegal activities and regulated goods - Weapons	0.179	0.1960	Figure represents all Illegal activities and regulated goods, which includes . 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	0.179	0.1960	Figure represents all Illegal activities and regulated goods, which includes . 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	0.081	0.0870	
Online piracy	Integrity and authenticity - Intellectual property violations	0.024	0.0120	Figure Represents all Integrity and Authenticity, including: . 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: Policy Violation Score

Percentage volume of videos removed by policy violation

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Hate speech & acts of aggression	Hateful behavior	**0.02**	**0.0080**	Figure represents all Hateful Behavior, which includes . 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	**0.02**	**0.0080**	Figure represents all Hateful Behavior, which includes . 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	**0.179**	**0.1960**	Figure represents all Illegal activities and regulated goods, which includes . 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	**0.024**	**0.0120**	Figure Represents all Integrity and Authenticity, including: . Online Piracy . Spam This figure also includes videos removed for Misinformation, as defined by TikTok Community Guidelines
Terrorism	Violent extremism	**0.003**	**0.0870**	
Debated Sensitive Social Issue	Hateful behavior	**0.02**	**0.0080**	Figure represents all Hateful Behavior, which includes . Debated Sensitive Social Issue . Hate speech & acts of aggression . Obscenity and Profanity

Question 2: How safe is the platform for advertisers?

Not submitted

<table>
<tr><th>GARM Category</th><th>Relevant Policy</th><th>Latest Period</th><th>Previous Period</th><th>Commentary</th></tr>
<tr><td rowspan="2"><b>Adult &amp; Explicit Sexual Content</b></td><td>Minor Safety - Sexual exploitation of minors</td><td rowspan="13" colspan="3">This is not something we currently track. However, content that appears either side of in-feed ads is moderated through AI and human reviewers.
Because our ads are 100% share of voice (full screen), there is 0% on screen adjacency</td></tr>
<tr><td>Adult nudity and sexual activities</td></tr>
<tr><td><b>Arms &amp; Ammunition</b></td><td>Illegal activities and regulated goods - Weapons</td></tr>
<tr><td><b>Crime &amp; Harmful acts to individuals and Society, Human Right Violations</b></td><td>Illegal activities and regulated goods - Criminal Activities</td></tr>
<tr><td><b>Death, Injury or Military Conflict</b></td><td>Violent and graphic content</td></tr>
<tr><td><b>Online piracy</b></td><td>Integrity and authenticity - Intellectual property violations</td></tr>
<tr><td><b>Hate speech &amp; acts of aggression</b></td><td>Hateful behavior</td></tr>
<tr><td><b>Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust</b></td><td>Hateful behavior – Slurs</td></tr>
<tr><td><b>Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol</b></td><td>Illegal activities and regulated goods - Drugs, controlled substances, alcohol, and tobacco</td></tr>
<tr><td><b>Spam or Harmful Content</b></td><td>Integrity and authenticity - Spam and fake engagement</td></tr>
<tr><td><b>Terrorism</b></td><td>Violent extremism</td></tr>
<tr><td><b>Debated Sensitive Social Issue</b></td><td>Hateful behavior</td></tr>
</table>

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

Authorized Metric: Automatic blocks of content

Percentage of violating views removed by technology methods

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	95.8%	Not applicable – not tracking for H1 2020	
	Adult nudity and sexual activities	90.6%		Minor Safety includes anything that may perpetuate the abuse, harm, endangerment, or exploitation of minors
Arms & Ammunition	Illegal activities and regulated goods - Weapons	94.8%		Figure represents all Illegal activities and regulated goods, which includes • 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	94.8%		Figure represents all Illegal activities and regulated goods, which includes • 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	92.7%		.
Online piracy	Integrity and authenticity - Intellectual property violations	91.3%		Figure Represents all Integrity and Authenticity, including: • 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?

Authorized Metric: Automatic blocks of content

Percentage of violating views removed by technology methods

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Hate speech & acts of aggression	Hateful behavior	83.5%	Not applicable – not tracking for H1 2020	Figure represents all Hateful Behavior, which includes • 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	83.5%		Figure represents all Hateful Behavior, which includes • 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	94.8%		Figure represents all Illegal activities and regulated goods, which includes • 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	91.3%		. Figure Represents all Integrity and Authenticity, including: • Online Piracy • Spam This figure also includes videos removed for Misinformation, as defined by TikTok Community Guidelines
Terrorism	Violent extremism	89.4%		
Debated Sensitive Social Issue	Hateful behavior	83.5%		Figure represents all Hateful Behavior, which includes • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity

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

Authorized Metric: Automatic blocks of content

Percentage of violating views removed by technology methods

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	97.1%	In H1 2020, we didn't have a figure broken down by category. Our overall figure was 96.4%	Minor Safety includes anything that may perpetuate the abuse, harm, endangerment, or exploitation of minors
	Adult nudity and sexual activities	88.3%		
Arms & Ammunition	Illegal activities and regulated goods - Weapons	96.3%		Figure represents all Illegal activities and regulated goods, which includes • 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	96.3%		Figure represents all Illegal activities and regulated goods, which includes • 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	93.2%		.
Online piracy	Integrity and authenticity - Intellectual property violations	70.5%		Figure Represents all Integrity and Authenticity, including: • 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?

Authorized Metric: Automatic blocks of content

Percentage of violating views removed by technology methods

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Hate speech & acts of aggression	Hateful behavior	**72.9%**	In H1 2020, we didn't have a figure broken down by category. Our overall figure was 96.4%	Figure represents all Hateful Behavior, which includes • 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	**72.9%**		Figure represents all Hateful Behavior, which includes • 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	**96.3%**		Figure represents all Illegal activities and regulated goods, which includes • 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	**70.5%**		. Figure Represents all Integrity and Authenticity, including: • Online Piracy • Spam This figure also includes videos removed for Misinformation, as defined by TikTok Community Guidelines
Terrorism	Violent extremism	**86.9%**		
Debated Sensitive Social Issue	Hateful behavior	**72.9%**		Figure represents all Hateful Behavior, which includes • Debated Sensitive Social Issue • Hate speech & acts of aggression • Obscenity and Profanity

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

Authorized Metric: Automatic blocks of content

Percentage of violating views removed by technology methods

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	0 views – 83.3% 1+ views – 16.7%	0 views – 90.5% 1+ views – 9.5%	Total Across All Guidelines Figure
	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

Content removed by TikTok and then appealed by users

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Minor Safety - Sexual exploitation of minors	We reinstated 2,927,391 videos after they were appealed	Not applicable – not tracking in H1 2020	Content reinstatement not available for H1 2020 For H2 2020, 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. 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 you won't find 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 example, we don't monetize search terms related to the coronavirus pandemic.

And then there's the more obvious stuff. 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. And we're streamlining our logging so that we can be more transparent about our efforts, starting with Q4 2020.

Because Pinterest is personal media—not social media—things are a little different around here. People use Pinterest to curate ideas for themselves and their own lives. That means there are two types of surfaces on Pinterest: discovery surfaces that are more "public," like the home feed, and more "personal" surfaces, like boards and profiles that "belong" to individual users. Here's the important part: your ad only shows up on discovery surfaces, including home feed, search, and related Pins. Unlike social media, where users broadcast their interests to others, Pinners may use these more "personal" surfaces for independent projects and private interests. One result is that we do detect and remove a lot of adult content on Pinterest.

But not all content on Pinterest shows up on the public surfaces where we show ads—for instance, adult content mostly stays in those "personal" non-monetized spaces until we are able to remove it. When our automated tools detect potential adult content, for example, we prevent it from appearing on public surfaces where we show ads, and where other users might discover it. If we determine that these Pins violate our policies, we remove them. Then we use automated tools to remove any other instances of that image from the rest of our platform. So while we detected and removed a lot of adult content this quarter, those Pins comprised only 2.1M distinct images. More importantly, not a lot of people saw it. In fact, **98% of adult content that was removed on Pinterest was seen by fewer than 100 people during the reporting period.**

Our content policies and moderation practices are always evolving. For example, during the US election season, we removed false and misleading content that might interfere with the election process, including conspiracy theories and any content that could impede someone's ability to vote. We also have longstanding efforts to identify and remove medical misinformation, such as anti-vaccine content and false or misleading information about COVID-19. And recently, we've expanded our tactics to fight spam with Guardian, a real-time analytics and rules engine that we created, which allowed us to r**educe spam prevalence in Q4 by 35%.**

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.

Question 1: How safe is the platform for consumers?

Not submitted

Current Period: Q4 2020

Prior Period: Due to developments in our logging practices, we are only reporting transparency metrics for Q4 2020 and beyond, as prior numbers are not directly comparable.

		Prevalence
GARM Category	**Pinterest Category**	**Q4 2020**
Adult and explicit sexual content	**Adult sexual services**	**OUT OF SCOPE**
	Adult content	**OUT OF SCOPE**
Crime & harmful acts to individuals & society; human rights violations	**Harassment & criticism**	**OUT OF SCOPE**
	Self-injury and harmful behavior	**OUT OF SCOPE**
Dangerous goods	**Dangerous goods and activities**	**OUT OF SCOPE**
Death, injury, military conflict	**Graphic violence and threats**	**OUT OF SCOPE**
Debated sensitive social issues	**Conspiracy theories**	**OUT OF SCOPE**
	Medical misinformation	**OUT OF SCOPE**
	Civic misinformation	**OUT OF SCOPE**
Hate speech	**Hateful activities**	**OUT OF SCOPE**
Spam and malware	**Spam**	**3.51%**
	Methodology notes	Of total impressions, the percentage that were of a Pin that was later removed as policy-violating. Based on a statistically significant sample of global users

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

Authorized Metric: Distinct Images Removed, Pins Removed, Boards Removed, Automatic Pin Removals

Content types removed by Pinterest - Content types automatically removed by Pinterest - Content removals by views

Current Period: Q4 2020

Prior Period: Due to developments in our logging practices, we are only reporting transparency metrics for Q4 2020 and beyond, as prior numbers are not directly comparable.

		Reach of Directly Deactivated Pins: Q4 2020								
GARM Category	**Pinterest Category**	**0 people**	**% 0 people by category**	**<10 people**	**% <10 by category**	**10-100 people**	**% 10-100 people by category**	**100+ people**	**% 100+ people by category**	**Total**
Adult and explicit sexual content	**Adult sexual services**	**76**	**10.6%**	**182**	**25.5%**	**203**	**28.4%**	**253**	**35.4%**	**100.0%**
	Adult content	**38,013,158**	**76.2%**	**8,570,147**	**17.2%**	**2,264,151**	**4.5%**	**1,008,225**	**2.0%**	**100.0%**
Crime & harmful acts to individuals & society; human rights violations	**Harassment & criticism**	**36,537**	**78.8%**	**7,421**	**16.0%**	**1,402**	**3.0%**	**1,011**	**2.2%**	**100.0%**
	Self-injury and harmful behavior	**159,202**	**90.7%**	**14,127**	**8.0%**	**1,408**	**0.8%**	**847**	**0.5%**	**100.0%**
Dangerous goods	**Dangerous goods and activities**	**21,354**	**50.5%**	**16,052**	**37.9%**	**2,925**	**6.9%**	**1,979**	**4.7%**	**100.0%**
Death, injury, military conflict	**Graphic violence and threats**	**1,887**	**50.3%**	**510**	**13.6%**	**559**	**14.9%**	**794**	**21.2%**	**100.0%**
Debated sensitive social issues	**Conspiracy theories**	**1,377,750**	**91.1%**	**116,929**	**7.7%**	**13,473**	**0.9%**	**4,069**	**0.3%**	**100.0%**
	Medical misinformation	**15,448**	**85.0%**	**1,418**	**7.8%**	**294**	**1.6%**	**1,024**	**5.6%**	**100.0%**
	Civic misinformation	**10,180**	**64.4%**	**4,300**	**27.2%**	**925**	**5.9%**	**404**	**2.6%**	**100.0%**
Hate speech	**Hateful activities**	**6,146**	**73.2%**	**902**	**10.7%**	**530**	**6.3%**	**819**	**9.8%**	**100.0%**
Spam and malware	**Spam**	**2,885,265**	**85.5%**	**175,592**	**5.2%**	**130,424**	**3.9%**	**183,888**	**5.4%**	**100.0%**
	Methodology notes	**Calculated based on the number of unique users in this reporting period that saw a policy-violating Pin for at least 1 second before it was directly removed.**								

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

Authorized Metric: Distinct Images Removed, Pins Removed, Boards Removed, Automatic Pin Removals

Content types removed by Pinterest - Content types automatically removed by Pinterest - Content removals by views

Current Period: Q4 2020

Prior Period: Due to developments in our logging practices, we are only reporting transparency metrics for Q4 2020 and beyond, as prior numbers are not directly comparable.

Removals + Reach: How effective is the platform in policy enforcement

		Number of Images Directly Removed[1]	Number of Pins Directly Removed[2]	Number of Boards Directly Removed[3]	Number of Accounts Directly Removed[4]	Number of Directly Removed Pins that were Automatically Removed	
GARM Category	**Pinterest Category**	**Q4 2020**	**Q4 2020**	**Q4 2020**	**Q4 2020**	**Q4 2020**	**Q4 2020 %**
Adult and explicit sexual content	**Adult sexual services**	**707**	**714**	**488**	**494**	**0**	**0.00%**
	Adult content	**2,100,253**	**49,855,681**	**50,767**	**7,754**	**1,565**	**0.003%**
Crime & harmful acts to individuals & society; human rights violations	**Harassment & criticism**	**3,763**	**46,371**	**795**	**816**	**43,756**	**94.36%**
	Self-injury and harmful behavior	**3,499**	**175,584**	**898**	**26**	**171,969**	**97.94%**
Dangerous goods	**Dangerous goods and activities**	**5,501**	**42,310**	**956**	**179**	**40,341**	**95.35%**
Death, injury, military conflict	**Graphic violence and threats**	**1,754**	**3,750**	**381**	**11**	**1,858**	**49.55%**
Debated sensitive social issues	**Conspiracy theories**	**52,863**	**1,512,221**	**1,877**	**219**	**489,679**	**32.38%**
	Medical misinformation	**5,938**	**18,184**	**345**	**13**	**11,564**	**63.59%**
	Civic misinformation	**3,238**	**15,809**	**456**	**24**	**542**	**3.43%**
Hate speech	**Hateful activities**	**1,980**	**8,397**	**4,604**	**2,487**	**5,665**	**67.46%**
Spam and malware	**Spam**	**1,378,472**	**3,375,169**	**2**	**3,115,438**	**3,375,169**	**100.00%**
	Methodology notes	1 Distinct images from Pins that are directly deactivated for violating policy, including those manually removed by an agent and those removed by match-detection tools as matching the images in manually deactivated Pins. Does not include distinct images that were removed because they were on a board that was deactivated or belonged to a user that was deactivated.	2 Pins that are directly deactivated for violating policy, including Pins manually removed by an agent and Pins removed by automated tools as matching manually deactivated images. Does not include Pins that were removed because they were on a board that was deactivated or belonged to a user that was deactivated.	3 Boards that are directly deactivated for violating policy. When policy-violating boards are removed, 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 Accounts that are directly deactivated for violating policy. When policy-violating accounts are removed, all Pins and boards belonging to those accounts are removed as well.	Pins that are directly deactivated without action from an agent. Note that Pins deactivated by match-detection tools as matching images manually deactivated are counted in column E and as manually deactivated content.	
	Commentary	Conspiracy theories: As part of our efforts to maintain an inspirational platform leading up to the US election in Nov 2020, we proactively removed new and pre-existing content that violated our conspiracy theory policy.				Adult content: While a smaller portion of Pins are automatically removed, our automated tools detect potential adult content and prevent it from appearing on public surfaces where we show ads or where other users might discover it. If we determine that these images violate our policies, we manually remove them.	

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

Authorized Metric: Account Appeals, Account Reinstatements

Accounts that are removed and then appealed by users, Accounts that have been reinstated after an appeal

Current Period: Q4 2020

Prior Period: Due to developments in our logging practices, we are only reporting transparency metrics for Q4 2020 and beyond, as prior numbers are not directly comparable.

		Account Appeals		Account Reinstatements
GARM Category	**Pinterest Category**	**Q4 2020**		**Q4 2020**
Adult and explicit sexual content	**Adult sexual services**	**38**		**2**
	Adult content	**1,813**		**355**
Crime & harmful acts to individuals & society; human rights violations	**Harassment & criticism**	**24**		**5**
	Self-injury and harmful behavior	**1**		**0**
Dangerous goods	**Dangerous goods and activities**	**6**		**0**
Death, injury, military conflict	**Graphic violence and threats**	**7**		**1**
Debated sensitive social issues	**Conspiracy theories**	**22**		**12**
	Medical misinformation	**3**		**0**
	Civic misinformation	**3**		**0**
Hate speech	**Hateful activities**	**31**		**7**
Spam and malware	**Spam**	**99,839**		**64,777**
	Methodology notes			

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.

For us, nothing is more important than the safety of Snapchat users and we have zero tolerance for using Snapchat for illicit purposes. We are as proactive as possible in detecting, preventing and acting on this type of abuse -- but we know bad actors are constantly evolving how they try to evade the rules on many platforms.

Snapping -- or talking with pictures -- was born out of our realization that the camera, which was once a tool for documenting important moments, could become a powerful platform for self-expression and visual communication. It's why Snapchat opens directly to the camera, and not a feed of content. It's why we made Snap's delete by default -- because until social media platforms came along, friends didn't keep a permanent transcript of every conversation they had. It's why Snapchat is centered around communication with a close network of people you actually know in real life, rather than a town square where anyone has the right to distribute anything to anyone without moderation.

Over the years, these design decisions have helped us protect our community from misinformation and toxicity. We use design development processes that consider the privacy, safety and ethical implications of a new feature at the front end of the process -- and don't launch it if it doesn't pass our intensive reviews.

We focus on making our features private-by-default, because just like in real life, we think individual users should choose what information they want to share and when. We don't offer an open newsfeed, instead we offer a content platform that is closed and only features news and entertainment from trusted media publishers and creators. We don't give anyone the opportunity to share unvetted content with a large audience on Snapchat and the majority of content on Snapchat is ephemeral -- all making it much harder for misinformation to 'go viral.'

We began publishing bi-annual Transparency Reports in 2015, offering important insight into the violating content we enforce against governmental requests for Snapchatters' account information and other legal notifications. In 2020, we began publishing content and account removal data points, and will continue to do so bi annually going forward. From 1 January 2020 – 30 June 2020, we enforced against 3,872,218 pieces of content, globally, for violations of our Community Guidelines -- **which amounts to less than 0.012% of all Story postings.**

Our commitment to building technology for humans, and not the other way around, has informed every product decision we have made. We take a human-centric approach to innovation, which means we start with a universal set of values at our core and deliberately consider people's needs ahead of only data. We have programs in place to evaluate the potential impact of a new feature on the safety, privacy and wellbeing of both Snapchatters our individual users and society during the product development process -- and if we think it will have a negative impact, it doesn't get released.

In collaboration with GARM and its members, we are committed to expanding on the data points and content categories in future reports, and aligning on ways to give our community a clear understanding of our safety policies and practices.

Question 1: How safe is the platform for consumers?

Authorized Metric: Policy Violation Rate

Percentage of Snaps removed as a percentage of total story postings

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Sexually Explicit Content	Content removed represented less than 0.012% of Story Postings	Content removed represented less than 0.012% of Story Postings	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: Policy Violation Rate

Percentage of Snaps removed as a percentage of total story postings

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Hate speech & acts of aggression	Hate Speech	Content removed represented less than 0.012% of Story Postings	Content removed represented less than 0.012% of Story Postings	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 2: How safe is the platform for advertisers?

Authorized Metric: Policy Violation Rate

Percentage of Snaps removed as a percentage of total story postings

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Adult & Explicit Sexual Content	Sexually Explicit Content	Content removed represented less than 0.012% of Story Postings	Content removed represented less than 0.012% of Story Postings	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: Policy Violation Rate

Percentage of Snaps removed as a percentage of total story postings

GARM Category	Relevant Policy	Latest Period	Previous Period	Commentary
Hate speech & acts of aggression	Hate Speech	Content removed represented less than 0.012% of Story Postings	Content removed represented less than 0.012% of Story Postings	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?

Authorized Metric: 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	3.119,948	1,160,881	2,930,946	747,797	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	234,527	137,721	248,581	140,583	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	183,929	141,314	246,629	176,912	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	183,929	141,314	246,629	176,912	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	104,523	59,131	63,917	34,574	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?

Authorized Metric: 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	31,041	26,857	46,936	41,381	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	234,527	137,721	248,581	140,583	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	104,523	59,131	63,917	34,574	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	N/A	<10	N/A	N/A	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?

Authorized Metric: Content Actioned, Actors Actioned

Content removed by Snap - Users removed by Snap

While we've been publishing Transparency Reports since 2015, we launched our first Transparency Report that featured content removal metrics in September 2020, reflective of 2H 2019. In the first two Transparency Reports featuring content removal, we focused on widening our category reporting and the inclusion of a prevalence metric as top priority. We are committed to adding proactive detection metrics for more categories in future reports, and did publish a proactive detection metric for CSAM in reporting period 1H 2020.

GARM Category	Relevant Policy	Latest Period	Previous	Commentary
Adult & Explicit Sexual Content	Child Sexual Exploitation and Abuse	**1H 2020**	**N/A**	Per Snap's 1H 2020 Transparency Report: We use PhotoDNA technology to proactively identify and report the uploading of known imagery of child sexual exploitation and abuse, and we report any instances to the authorities. Of the total accounts enforced against for Community Guidelines violations, we removed 2.99% for CSAM takedown. Moreover, Snap proactively deleted 70% of these.
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 3: How Effective is the Platform in Enforcing Safety Policy?

Authorized Metric: 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						

Appendices & FAQ

How is the report created and what is the governance?

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

STEP 1: Platforms involved in GARM confirm participation

STEP 2: GARM Working Group distributes data submission and commentary submission template

STEP 3: WFA aggregates submissions and GARM Steer Team develops analysis for Executive Summary

STEP 4: GARM platforms review and confirm content for accuracy and GARM Working Group approves content

STEP 5: WFA GARM publishes report

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

Why are we focusing on these four core questions?

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

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

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

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 1:

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**	**Not Submitted**	**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**	**Not Submitted**	**Not Submitted**
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**
	Removal of violating accounts by views	**Authorized Metric**	**Not Submitted**	**Not Submitted**	**Next Best Measure**	**Authorized Metric**	**Not Submitted**	**Authorized Metric**
	Removal of violating accounts	**Authorized Metric**	**Authorized Metric**	**Not Submitted**	**Authorized Metric**	**Not Submitted**	**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**	**Not Submitted**	**Not Submitted**
	Reinstatements (pieces of content)	**Authorized Metric**	**Authorized Metric**	**Authorized Metric**	**Not Submitted**	**Not Submitted**	**Not Submitted**	**Not Submitted**

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 Content Category	Relevant Platform Policy						
	YouTube	Facebook	Instagram	Twitter	TikTok	Pinterest	Snap
Adult & Explicit Sexual Content	Nudity & Sexual Content Child Safety	Adult Nudity and Sexual Activity	Adult Nudity and Sexual Activity	Non-Consensual Nudity Sensitive Media Child Sexual Exploitation	Minor safety – sexual exploitation of minors Adult nudity and sexual activities	Adult Sexual Services Adult Content	Sexually Explicit Content
Arms & Ammunition	Firearms	Regulated Goods: Firearms	Regulated Goods: Firearms	Illegal or certain regulated good or services	Illegal activities and regulated goods - weapons	Dangerous Goods and Activities	Regulated Goods
Crime & Harmful acts to individuals and Society, Human Right Violations	Harmful or Dangerous Content Hate Speech Harassment or cyberbullying	Violent and Graphic Content Bullying and Harassment Child Nudity and Sexual Exploitation Suicide and Self-Injury	Violent and Graphic Content Bullying and Harassment Child Nudity and Sexual Exploitation Suicide and Self-Injury	Violence Abuse and harassment	Illegal activities and regulated goods -criminal activities	Child Sexual Exploitation Self-Harm Harassment & Criticism	Threatening / Violence / Harm:
Death, Injury or Military Conflict	Violent or Graphic Content Harmful or Dangerous Content Suicide & Self-Injury	Violent and Graphic content	Violent and Graphic content	Promoting Self-harm	Violent and Graphic Content	Graphic Violence and Threats	Threatening / Violence / Harm
Online piracy	Fake Engagement Impersonation Sale of illegal or regulated goods or services YouTube Terms of Service	Intellectual Property Copyright Intellectual Property Counterfeit Intellectual Property Trademark	Intellectual Property Copyright Intellectual Property Counterfeit Intellectual Property Trademark	Copyright Trademark	Integrity and authenticity – intellectual property violations	Copyright Trademark	Spam
Hate speech & acts of aggression	Hate Speech	Hate speech Bullying and Harassment	Hate speech Bullying and Harassment	Hateful Conduct	Hate Speech Hateful Behavior	Hateful Activities	Threatening / Violence / Harm
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	Violent or Graphic Content Age Restriction	Hate Speech Bullying and Harassment	Hate Speech Bullying and Harassment	Sensitive Media	Hateful Behavior – Slurs Harassment & Bullying	Harassment & Criticism	
Illegal drugs, tobacco, e-cigarettes, vaping	Sale of Illegal or Regulated Goods or Services Harmful or dangerous content	Regulate Goods: Drugs	Regulate Goods: Drugs	Illegal or certain regulated goods or services	Illegal activities and regulated goods – drugs, controlled substances, alcohol and tobacco	Dangerous Goods and Activities	Regulated Goods
Spam & Malware	Spam, Deceptive Practices & Scams	Spam	Spam	Private Information Impersonation Platform manipulation	Integrity and authenticity – spam and fake engagement	Spam	Spam
Terrorism	Violent criminal organizations	Dangerous Organizations: Terrorism Dangerous Organizations: Organized Hate	Dangerous Organizations: Terrorism Dangerous Organizations: Organized Hate	Terrorism or Violent Extremism	Violent Extremism Dangerous individuals and organizations - Terrorists and terrorist organizations	Violent Actors	Terrorism
Debated Sensitive Social Issues		Hate Speech Bullying and Harassment	Hate Speech Bullying and Harassment		Hateful Behavior	Civic Misinformation Conspiracy Theories Medical Misinformation	
Other	COVID Misinformation Policy	COVID-19 and Vaccine Policy and Protections	COVID-19 and Vaccine Policy and Protections	Covid Integrity Covid-19 Misleading Information			

WFA

World Federation of Advertisers
London, Brussels, Singapore, New York

wfanet.org

info@wfanet.org

+32 2 502 57 40

twitter **@wfamarketers**
youtube.com/**wfamarketers**
linkedin.com/**company/wfa**