Aotearoa New Zealand Code of Practice for Online Safety and Harms

Report Template

Signatory:	**Google**

***If applicable*:** **Relevant Products / Services:**	**YouTube**

4.1 Reduce the prevalence of harmful content online

Signatories commit to implementing policies, processes, products and/or programs that would promote safety and mitigate risks that may arise from the propagation of harmful content online, as it relates to the themes identified in section 1.4.

Outcome 1. Provide safeguards to reduce the risk of harm arising from online **child sexual exploitation & abuse (CSEA)**

Background: How YouTube Moderates Content

There are inherent tensions that come with fulfilling our mission to organize the world's information and make it universally accessible and useful. We must strike a careful balance between the free flow of information, safety, efficiency, accuracy, and other competing values.

YouTube is an open video platform where anyone can upload a video and share it with the world. With this openness comes incredible opportunities, as well as challenges – which is why we're always working to balance creative expression with our responsibility to protect the community from harmful content. At the heart of our approach are the four Rs principles - four complementary levers to support information quality and moderate content:

- We Remove content that violates our Community Guidelines. YouTube's Community Guidelines set out clear categories of content that are prohibited on our platform. These policies apply to content on our platform including videos, video descriptions, comments, live streams, links and any other YouTube products or features. We constantly evaluate our policies and enforcement guidelines and will continue to consult with experts and the community and make changes as needed. We take steps to ensure our policies evolve to keep pace with emerging challenges.

For content that does not violate our Community Guidelines but runs up against local laws, we will assess the requests to block this content on legal grounds. This may include content classified as ‘objectionable’ under New Zealand’s Films, Videos, Publications and Classifications Act.

- We Reduce the spread of harmful misinformation and content that brushes up against our policy lines. In 2019, we announced changes to our recommendation systems to reduce the spread of borderline content, resulting in a 70% drop in watch time on non-subscribed, recommended content in the US that year. We also saw a drop in watch time of borderline content coming from recommendations in other markets. And as of March 2021, we rolled out changes to our recommendation system to reduce borderline content in every market where we operate. We are committed to continuing our work to reduce recommendations of borderline content. While algorithmic changes take time to ramp up and consumption of borderline content might go up and down, our goal is to keep views of non-subscribed, recommended borderline content below 0.5% of overall views.

- We Raise up authoritative sources when people are looking for news and information through a range of features including through raising authoritative voices for newsworthy events and topics prone to misinformation, the introduction of product features such as the Breaking News shelf which feature relevant videos from authoritative news sources, and by providing context to enable users to evaluate information through information panels that feature text-based information alongside certain search results and videos.

- We Reward trusted, eligible creators and artists through the YouTube Partner Programme (YPP), which enables creators to start monetising their content, as well as gaining access to dedicated support and benefits. Over the last few years, we have taken steps to strengthen the requirements for monetisation so that spammers, impersonators and other offenders can't hurt the ecosystem or take advantage of creators who have put their time, energy and passion into producing high-quality content. To apply for membership of YPP, channels must meet eligibility thresholds related to watch time and subscribers. After they have applied, YouTube's review team ensures that only channels that meet eligibility thresholds and follow all of our guidelines are admitted to the programme, which makes them eligible to receive access to ads and other monetisation products. Advertisers typically do not want to be associated with controversial or sensitive content on YouTube, as defined in our advertiser-friendly content guidelines. If a creator has turned on ads monetisation for a video but our reviewers and automated systems determine that the video does not comply with our advertiser-friendly content guidelines, the video will have limited or no ads appear against it, which means that they won't be able to make money on that video. We may also suspend a creator's channel from the YPP for severe or repeated violations of our YouTube monetisation policies.

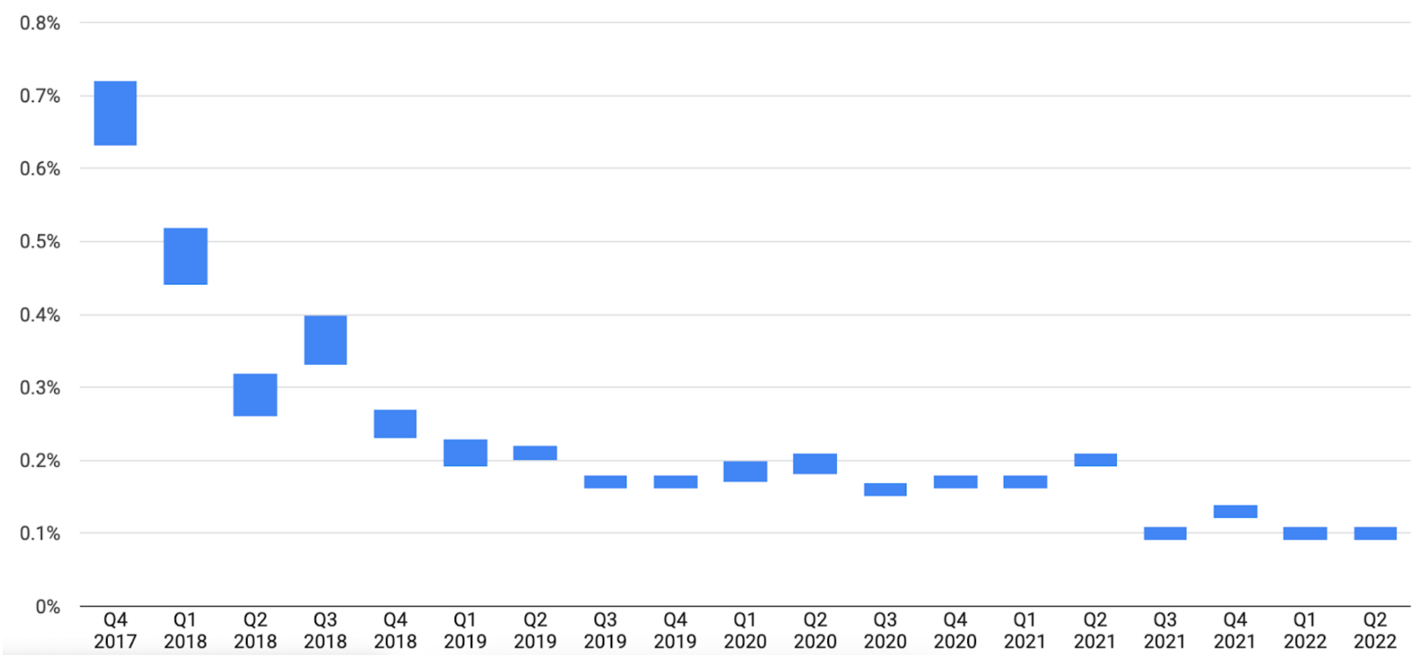
We enforce our policies at scale through a combination of people and machine learning. For example, we sometimes use hashes (or “digital fingerprints”) to automatically identify copies of known violative content before they are made available for viewing. The YouTube community also plays an important role in flagging videos that violate our Community Guidelines. Any logged-in user can flag a video by clicking on the three dots to the bottom right of the video player and selecting “Report”. Users can also report inappropriate channels, playlists, comments and other content.

Additionally, we have developed the Trusted Flagger program, which provides robust content reporting processes for individuals, government agencies, and non-governmental organizations (NGOs) that are particularly effective at notifying YouTube of content that violates our Community Guidelines. Participants of the Program receive a policy training and have a dedicated path of communication with our Trust & Safety team. Because participants' flags have a higher action rate than the average user, we prioritize them for review.

Trained teams evaluate a video before taking action to ensure it actually violates our policies and to protect content that has an educational, documentary, scientific, or artistic purpose. These teams are located in countries around the world, are fluent in multiple languages, and carefully evaluate flags 24 hours a day, seven days a week. Reviewers have extensive training in YouTube's Community Guidelines and often specialize in specific policy areas such as child safety or hate speech. They remove content that violates our terms, age-restrict content that may not be appropriate for all audiences, and leave content live when it doesn't violate our guidelines. Reviewers' inputs are used to train and improve our machine learning systems. There are also limited cases where we automatically remove illegal content, such as child sexual abuse material (CSAM).

YouTube strives to remove content that violates our Community Guidelines before users are exposed to this content. To measure our progress on removing violative videos, we have developed a metric called Violative View Rate (VVR). This metric is an estimate of the proportion of video views that violate our Community Guidelines in a given quarter (excluding spam). In order to calculate VVR, we take a sample of the views on YouTube and send the sampled videos for review. Once we receive the decisions from reviewers about which videos in the sample are violative, we aggregate these decisions in order to arrive at our estimate. Additional data on VVR can be found here. A more detailed explanation of the VVR calculation (including which policies are included) is found within the Community Guidelines Report FAQs, and further information on the VVR methodology can be found here.

In **Q2** (Apr - Jun 2022), VVR was **0.09 - 0.11%** (i.e., out of every 10,000 views on YouTube,
9-11 were of violative content).



Where content is found to violate our Community Guidelines, we remove the content and send a notice to the creator. The first time that a creator violates our Community Guidelines, they typically receive a warning with no penalty to the channel. After one warning, we'll issue a Community Guidelines strike to the channel and the account will have temporary restrictions including not being allowed to upload videos, live streams or stories for a one-week period. Channels that receive three strikes within a 90-day period will be terminated. Channels that are dedicated to violating our policies or that have a single case of severe abuse of the platform will bypass our strikes system and be terminated.

All strikes and terminations can be appealed if the creator believes that there was an error, and our teams will re-review the decision (more information is available here). There is more information about our strikes system here and our channel or account terminations here.

We provide regular, publicly available reports on enforcement of our content policies. Further detail is provided in response to Outcome 10 below. The following is a list of these reports (including public links to the materials):

- Our YouTube Community Guidelines Enforcement report provides a quarterly update on the work we do to enforce our policies on YouTube. The report offers data on global video, channel, and comment removals for violating our policies; appeals and reinstatements; and human and machine flagging.
- Our Google Transparency Report website is a centralised hub for transparency reporting on key topics including child safety, copyright, and government requests to remove content.
- The annual Ads Safety Report provides updates on policy enforcement in Google Ads.

- Our Threat Analysis Group Quarterly Bulletin (published on our Threat Analysis Group blog) discloses actions we have taken against coordinated influence operation campaigns on our platforms.

Outcome 1: Child Sexual Exploitation and Abuse

Google is committed to fighting child sexual exploitation and abuse online and preventing our platforms from being used to spread child sexual abuse material (CSAM). YouTube's Community Guidelines prohibit sexually explicit content featuring minors and content that sexually exploits minors.

- Our child safety policy prohibits content that endangers the emotional and physical well-being of minors. This includes content that sexually exploits minors, that highlights harmful or dangerous acts involving minors, and that causes emotional distress to minors.
- More generally it is important to note that explicit content that is meant to be sexually gratifying is prohibited on YouTube.

We devote significant resources - technology, people and time - to deterring, detecting, removing and reporting child sexual exploitation content and behaviour.

- We have heavily invested in engineering resources to detect CSAM in ways that are precise and effective, and have long used this technology to prevent the distribution of CSAM on YouTube.
    - Our proprietary CSAI Match technology allows us to detect and remove content previously identified as CSAM. Once we have identified a video as illegal and reported it to the National Center for Missing and Exploited Children (NCMEC), the content is hashed—i.e. given a unique digital fingerprint—and used to detect matching content. This hashing and scanning technology is highly precise at detecting known CSAM and enables us to detect illegal content more quickly. We maintain a database of known hashes and any content that is matched against this list is removed and reported to NCMEC.
    - We also use machine learning technology to identify CSAM that hasn't previously been identified.
    - In addition to our long-standing efforts to combat CSAM, we have made large investments to detect and remove content which may not meet the legal definition of CSAM, but where minors are still being sexualized or exploited.
- In the context of child safety, global partners under our Trusted Flagger program include the National Center for Missing and Exploited Children, Family Online Safety Institute, SaferNet, Childline South Africa, and ECPAT Indonesia.

We want to protect children using our products from experiencing grooming, sextortion, trafficking and other forms of child sexual exploitation. We continue to invest resources to ensure children and families have a safe experience and build products designed for children and families with high standards of privacy and safety protections.

- Users must be at least 13 years or over to create an account for themselves on our main YouTube service or a parent or legal guardian must enable it for them. YouTube employs machine learning to identify signals on YouTube channels that indicate when an account operating a channel may be owned by a user under 13. We rely on signals to find these channels, and then flag for a team to review more closely. If we suspect that an account has been created by someone under 13, we require them to verify their age with a credit card or a Government issued ID. If they cannot verify that they are above the age of 13, then they will be required to either add parental supervision or delete their account. We took action on more than 7M accounts globally during the first 3 quarters of 2021 when we learned they may belong to a user under the age of 13 – 3M of those in Q3 alone as we have ramped up our automated removal efforts.

- For users under 13, we have developed products to help provide children and families with more contained, age-appropriate experiences and protections and develop the balance that works for them.
    - YouTube Kids is an app that provides a separate YouTube experience designed especially for children, which parents can supervise. The app uses a mix of filters, content moderation, and user feedback to help keep the videos in YouTube Kids family-friendly, allowing children to explore a catalog of content in a safer environment. YouTube Kids can be used as a standalone app or web experience, where parents can sign-in to access a broader set of controls to customise their child's experience, or through Family Link. Content has to be age-appropriate and meet our quality principles before they can be made available on YouTube Kids.
    - Parents using Family Link can also choose to allow their children under the age of 13 to access the main YouTube service by selecting one of three content settings. The new YouTube Supervised Experience looks much like YouTube's flagship app and website, but with adjustments to the features children can use and ads protections. For example, on Supervised Experiences, we don't serve personalised ads and we prohibit certain types of advertising, including ads related to weight loss and diets or ads for dating sites. YouTube supervised experiences also have disabled in-app purchases, as well as features such as uploading videos or livestreams and reading or writing comments, to help prevent risks of unwanted interaction with other users.
- We age-restrict content that does not violate our Community Guidelines but may still not be appropriate for viewers under 18. Age-restricted content is not viewable by users below 18 years of age or who are signed out. The categories of content we consider for age-restriction include harmful or dangerous activities, nudity and sexually suggestive content, violent or graphic content, and vulgar language, videos that may contain adults participating in dangerous activities that minors could easily imitate, videos that invite sexual activity – such as provocative dancing – or videos with heavy profanity. We provide additional details about the types of content we consider for age-restriction here.

We have developed materials to guide parents in the support they can provide to their children. For example, YouTube recently published the guide "Exploring YouTube Confidently: A family guide to supervised experiences." The guide helps parents better understand YouTube Supervised Experience and the controls they can use, and provides helpful tips on how to talk to children about the content they watch, the time they spend online and how they can use the privacy controls we offer. YouTube also publishes videos to increase parental awareness about the supervision tools we offer.

We collaborate with the NCMEC and other organisations globally in our efforts to combat online child sexual abuse. As part of these efforts, we establish strong partnerships with NGOs and industry coalitions to help grow and contribute to our joint understanding of the evolving nature of child sexual abuse and exploitation.

- YouTube makes its CSAI Match technology available to partners in industry and NGOs. We give access to fingerprinting software and an API to identify matches against our database of known abusive content.
- In addition to reporting identified CSAM to the NCMEC, which liaises with law enforcement agencies around the world, we also contribute to the NCMEC hash database. These hashes help other platforms identify CSAM at scale. Contributing to the NCMEC hash database is one of the important ways that we, and others in the industry, can help in the effort to combat CSAM because it helps reduce the recirculation of this material and the associated re-victimisation of children who have been abused.
- Google is an active member of several coalitions, such as the Technology Coalition, the ICT Coalition, the WeProtect Global Alliance, INHOPE and the Fair Play Alliance, that bring companies and NGOs together to develop solutions that disrupt the exchange of CSAM online and prevent the sexual exploitation of children. Together we fund child safety research and share tools and knowledge, such as our insights into transparency reporting, in-product detection and operational processes.
- Google is also a member of the Digital Trust and Safety Partnership (DTSP) - see Outcome 12 below.

Data on YouTube's efforts to combat online CSAM is included in Google's Transparency Report. Between July 2021 and December 2021, YouTube:

- made 123,963 CyberTipline reports to NCMEC; and
- reported 135,517 pieces of content to the NCMEC.

Over the same period, YouTube removed over 95,000 channels and over 3 million videos for violation of our child safety policies.

Outcome 2: Provide safeguards to reduce the risk of harm arising from online **bullying or harassment**

YouTube's mission is to give everyone a voice and show them the world. We believe that everyone deserves to have a voice, and that the world is a better place when we listen, share and build community through our stories. Harassment hurts our community by making people less inclined to share their opinions and engage with each other.

YouTube's Community Guidelines prohibit content that violates our harassment and cyberbullying policies. This policy protects specific individuals. We consider content harassment when it targets an individual with prolonged or malicious insults based on intrinsic attributes, including their protected group status or physical traits. This also includes harmful behavior such as threats, bullying, doxxing, or encouraging abusive fan behavior. More details about content that violates our guidelines can be found on our harassment and cyberbullying page.

We also recognise that harassment sometimes occurs through a pattern of repeated behaviour across multiple videos or comments, even when individual videos may not cross our policy line. Examples include:

- Repeatedly encouraging abusive audience behavior.
- Repeatedly targeting, insulting and abusing an identifiable individual based on their intrinsic attributes

across several uploads.

- Exposing an individual to risks of physical harm based on the local social or political context.
- Creating content that harms the YouTube community by persistently inciting hostility between creators for personal financial gain.

We remove content that violates these policies and issue strikes to creators for those violations according to our strikes policy. Channels in the YouTube Partner Programme (YPP) may also be suspended on review for repeat violations, eliminating their ability to make money on YouTube. This is in line with our goals to ensure that we reward only trusted creators on our platform.

Harassment is a complex policy area to enforce at scale, as decisions require nuanced understanding of local languages and contexts. To help us consistently enforce our policies, we have review teams with linguistic and subject matter expertise.

Between April and June 2022, YouTube removed 47,297 channels for harassment and cyberbullying (representing 1.2% of the 3,987,509 channels removed) and 11.1% of the 4,496,933 videos removed were for violations of our harassment and cyberbullying policies. See our YouTube Community Guidelines Transparency Report for further data.

Empowering our community

To help ensure that our community feels safe while using YouTube, we have a number of tools to protect our creators, artists and users:

- Anonymous reporting (flagging) of inappropriate or abusive content or users. YouTube integrates flagging directly into its services, so that users can easily see how to report online bullying or harassment.
- Moderation tools for comments so that creators can shape the tone of the conversation on their channels. We hold potentially inappropriate comments for review, so that creators can best decide what is appropriate for their audience. We also have other tools that empower creators to block certain words in comments, block certain individuals from commenting or assign moderation privileges to other people so that they can more efficiently monitor comments on their channel.
- To ensure that YouTube promotes respectful interactions between views and creators, we have a feature that will warn users if their comment might seem offensive to others, giving them the option to reflect and edit before posting.
- We have also made the ‘dislike' count private across YouTube. In early 2021, we experimented with the dislike button to see whether or not changes could help better protect our creators from harassment, and reduce dislike attacks - where people work to drive up the number of dislikes on a creator's videos.   As part of this experiment, viewers could still see and use the dislike button. But because the count was not visible to them, we found that they were less likely to target a video's dislike button to drive up the count. In short, our experiment data showed a reduction in dislike attacking behaviour. We also heard directly from smaller creators and those just getting started that they are unfairly targeted by this behaviour — and our experiment confirmed that this does occur at a higher proportion on smaller channels. We also want to ensure that creators and users have the information they need to stay safe on YouTube:

- Users can access our Community Guidelines and policies and instructions on how to report violations.
- We maintain a list of resources to help support users in feeling safe while using YouTube.
- We have established a Creator Safety Center with tips from creators, security experts and non-profit organisations to help creators make a plan to stay safe online.
- Our blog posts provide further detail on how YouTube works, including detail on our policies and updates.
- We run marketing campaigns to promote our online safety and security tools. For example, in New Zealand in 2022 we:
    - Published a two-page summary about how YouTube fosters child safety.
    - Delivered a Safer with Google campaign to promote online safety and security, raising awareness of safe browsing. Along with leveraging Google platforms, we shared information in newspapers and banners during August - November 2023.

Google

Setting digital ground rules can be tough

Google Family Link helps parents manage what's right for their family online

We're working directly with experts and educators to help families use technology in a way that's right for them. The Google Family Link app helps parents manage the apps their child downloads or set a device bedtime.

Explore how we're helping parents keep their kids safer online at safety.google/families

Safer with Google

Example from Safer with Google Campaign

YouTube is committed to education and awareness raising among its creators and users on how to prevent or stop online bullying and harassment. We also support a wide range of global and local programs including:

- Sponsorship of Netsafe's work that includes programs to educate and raise awareness to stop online bullying and harassment.
- Google's Be Internet Awesome, a global programme which aims to teach the fundamentals of online safety for children.

- The Digital Licence, developed by the Alannah & Madeline Foundation, a not-for-profit dedicated to keeping children safe from violence and bullying. The Digital Licence is an interactive online quiz providing cyber safety for kids; educating them on what to do if they are exposed to unwanted, inappropriate and offensive content or cyber bullying; and the consequences of putting their privacy at risk when interacting online. Google made this program available free of charge to all year 8 and 9 students in New Zealand.

Outcome 3: Provide safeguards to reduce the risk of harm arising from online **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, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their family members, and veteran status. This means we don't allow content that dehumanizes individuals or groups with these attributes, claims they are physically or mentally inferior, or praises or glorifies violence against them. We also don't allow use of stereotypes that incite or promote hatred based on these attributes, or racial, ethnic, religious, or other slurs where the primary purpose is to promote hatred.

In June 2019 we announced an update to our hate speech policy to specifically prohibit videos alleging that a group is superior in order to justify discrimination, segregation or exclusion based on attributes like age, gender, race, caste, religion, sexual orientation or veteran status. We also announced that we will remove content denying that well-documented violent events took place. We continue to consult with experts in subjects like violent extremism, supremacism, civil rights, and free speech from across the political spectrum for insight on emerging trends and continuously review our policies and make updates where appropriate.

YouTube is a platform for free expression, and this can be a delicate balancing act with the need to protect our community. In enforcing our hate speech policy, we consider the purpose of the video. We may allow content that includes discussion around hate speech if the purpose is educational, documentary, scientific, or artistic (EDSA) in nature. If users are posting content related to hate speech for this purpose, we encourage them to be mindful to provide enough information so viewers understand the context, such as through an introduction, voiceover commentary, or text overlays, as well as through a clear title and description. Providing documentary or educational context can help the viewer, and our reviewers, understand why potentially disturbing content sometimes remains live on YouTube.

More details about content that violates our guidelines can be found on our hate speech policy page.

If content directed against an individual is not covered by our hate speech policy, it may instead be covered by our policies against harassment and violence, while content that praises or glorifies terrorist or criminal figures or organizations is covered by our policies against violent criminal organizations. Reviewers evaluate flagged content against all of our Community Guidelines and policies.

Like harassment, hate speech is a complex policy area to enforce at scale, as decisions require nuanced understanding of local languages and contexts. To help us consistently enforce our policy, we have expanded our review team's linguistic and subject matter expertise. We're also deploying machine learning to better detect potentially hateful content to send for human review, applying lessons from our enforcement against other types of content, like violent extremism. Our Trusted Flagger program includes global and local partners like Faith Matters, JFDA, Licra, and Observatorio Web.

Between April and June 2022, YouTube removed 32,346 channels for being hateful or abusive (representing 0.8% of the 3,987,509 channels removed) and 145,688 videos (or 3.2% of the 4,496,933 videos removed).

See our YouTube Community Guidelines Transparency Report for further data.

In addition to removing content that violates our policies (and imposing strikes or account suspensions and terminations) we also have long-standing advertiser-friendly guidelines that prohibit ads from running on videos that include hateful content.

Channels that repeatedly violate our hate speech policies are suspended from the YouTube Partner program (YPP), meaning they can't run ads on their channel or use other monetization features, like Super Chat.

YouTube's products and tools that seek to prohibit or reduce hate speech (Measure 11) and our tools for users to report potential hate speech (Measure 12) are outlined under Outcome 2 above.

While YouTube seeks to tackle hate speech on the platform by enforcing our Community Guidelines, we also want to support users in thinking critically about the content that they see on YouTube and the online world so that they can make their own informed decisions on how to reduce or stop the spread of hate speech. We therefore welcome the opportunity to sponsor Netsafe's work, including programs to educate and raise awareness to stop online hate speech. This builds on global programs such as Google's Be Internet Awesome, detailed above, and provides support for Netsafe's Online Safety Week.

Outcome 4: Provide safeguards to reduce the risk of harm arising from online **incitement of violence**

Content encouraging others to commit violent acts is not permitted on YouTube.

- Our violent and graphic content policies prohibit content that incites others to commit violent acts against individuals or a defined group of people.
- Content intended to praise, promote or aid violent criminal organisations is prohibited under our violent criminal organisations policy. This includes content produced by violent criminal or terrorist organisations, content praising or memorialising prominent terrorist or criminal figures in order to encourage others to carry out acts of violence and content aimed at recruiting new members to violent criminal or terrorist organisations.
- We are also expanding our policies to combat violent extremism by removing content glorifying violent acts for the purpose of inspiring others to commit harm, fundraise, or recruit, even if the creators of such content are not related to a designated terrorist group.
- We do not allow content that includes instructions to kill or harm or which promotes or glorifies violent tragedies under our harmful or dangerous content policies.
- Our hate speech policy prohibits the incitement of violence against individuals or groups based on protected attributes (see Outcome 3 above).

Reviewers evaluate flagged content against all of our Community Guidelines and policies. We may make exceptions for content that has a clear educational, documentary, scientific, or artistic purpose. For example, we may allow content depicting terrorist violence published by major news outlets, but if the content is graphic, we may place it behind an interstitial warning users that the content is graphic in nature (e.g., documentary footage from a war zone). We think this policy is critical to striking the right balance with free expression, and serves important societal purposes. Note that we do not allow the following kinds of content even if there's educational, documentary, scientific, or artistic context provided:

- Violent physical sexual assaults (video, still imagery, or audio).
- Footage filmed by the perpetrator during a deadly or major violent event, in which weapons, violence, or injured victims are visible or audible.

Between April 2022 and June 2022, 72,990 (or 1.6%) of videos removed by YouTube were removed for promoting violence and violent extremism. Of the 3,987,509 channels terminated during this same period, 9,814 or 0.2% were removed on this same ground.

At YouTube we use a combination of smart detection technology and human reviewers that helps us quickly detect, review, and remove this type of content.

- For novel or new content, we have trained automated systems that identify content based on a wide range of signals, including imagery, words, and iconography. In the April - June quarter of 2022, we removed 4,496,933 videos from YouTube for violating our community guidelines.
    - 4,195,734 of these videos were first flagged by machines rather than humans.
    - Of those removed videos that were detected by machines, 72.4% received 10 views or less.
- Our Trusted Flagger program includes the International Center for the Study of Radicalization at King's College London, the Institute for Strategic Dialogue, the Wahid Institute in Indonesia, and government agencies focused on counterterrorism.

As a signatory to the Christchurch Call, we have committed to continuing to invest in technology that improves our capability to detect and remove terrorist and violent extremist content online, including the extension or development of digital fingerprinting and AI-based technology solutions.

In this context (and relevant also to Outcome 5), any channel that has received strikes against YouTube's Community Guidelines is temporarily prevented from uploading or live-streaming. YouTube has the following additional checks in place for live streaming:

- Channels wishing to live stream must verify their account via phone must wait 24 hours.
- We require that creators have at least 50 subscribers before they can post a live stream from a mobile or other portable device, such as a GoPro.
- Channels must have at least 1,000 subscribers and 4,000 public watch hours before they are eligible to apply to YouTube's Partner Program in order to monetise their content (including live streams).

YouTube's efforts in relation to Measures 17 and 18 are outlined under Outcomes 2 and 3 above.

In 2017, YouTube, Facebook, Microsoft, and Twitter founded the Global Internet Forum to Counter Terrorism (GIFCT) as a group of companies dedicated to disrupting terrorist abuse of members' digital platforms. Although our companies have been sharing best practices around counterterrorism for several years, GIFCT provided a more formal structure to accelerate and strengthen this work and present a united front against the online dissemination of terrorist content. In collaboration with the Tech Against Terrorism initiative, we have held workshops with more than 100 smaller tech companies around the world.

YouTube and GIFCT's other founding members signed on to the Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online. Building on the Christchurch Call, GIFCT developed a new content incident protocol for GIFCT member companies to respond efficiently to perpetrator-created content after a violent attack. This protocol has been tested and proven effective in allowing platforms to share real time situational awareness and hashes - for example following the attacks in Halle, Germany (October 2019); Glendale, Arizona, USA (May 2020); Buffalo, New York, USA (May 2022); and Memphis, Tennessee, USA (September 2022).

GIFCT has evolved to be a standalone organization with an independent Executive Director and staff. GIFCT's structure also includes an Independent Advisory Committee composed of government representatives and civil society members, including advocacy groups, human rights specialists, researchers and technical experts. Within the new governance framework of the institution, we will take a position on the independent GIFCT's Operating Board.

We have also invested in programs and projects to support initiatives that promote tolerance and understanding.

- Jigsaw, a unit within Google, developed the Redirect Method – an open-source methodology that uses targeted advertising to connect people searching online for harmful content with constructive alternative messages. Piloted by Jigsaw and Moonshot CVE in 2016 and subsequently deployed internationally by Moonshot in partnership with tech companies, governments and grassroots organizations, it uses pre-existing content made by communities across the globe, including content not created for the explicit purpose of countering harm, to challenge narratives which support violent extremism, violent misogyny, disinformation and other online harms. For example, it has been used to redirect white-supremacy and Neo-Nazi related search-terms to disengagement NGOs.

- Jigsaw has also partnered with leading experts at American University's Polarization and Extremism Research Innovation Lab (PERIL) to explore technological approaches to addressing online radicalization.

- We have supported Sticks & Stones in their efforts to reduce harmful content in digital spaces by adopting a mental health and community approach through a $1.4M Google.org grant.

Outcome 5: Provide safeguards to reduce the risk of harm arising from online **violent or graphic content**

Gory and violent content intended to shock or disgust viewers is not allowed on our platform.

- Our violent or graphic content policies prohibits violent and graphic content as well as animal abuse content and extends to dramatised or fictional content where the viewer is not given enough content to understand that the footage is dramatised or fictional.

Violent and graphic content may also be captured under our violent criminal organisations policy and harmful or dangerous content policies.

Reviewers evaluate flagged content against all of our Community Guidelines and policies. We may make exceptions for content that has a clear educational, documentary, scientific, or artistic purpose. For example, we may allow content depicting terrorist violence published by major news outlets but if the content is graphic, we may place it behind an interstitial warning users that the content is graphic in nature (e.g., documentary footage from a war zone). We think this policy is critical to striking the right balance with free expression, and serves important societal purposes. Note that we do not allow the following kinds of content even if there's educational, documentary, scientific, or artistic context provided:

- Violent physical sexual assaults (video, still imagery, or audio).
- Footage filmed by the perpetrator during a deadly or major violent event, in which weapons, violence, or injured victims are visible or audible.

YouTube’s automated flagging tools, detailed above, help us to detect, review and remove violent or graphic content. Between April 2022 and June 2022:

- 11,587 channels were removed for violating our violent or graphic content policies. This represented 0.3% of the total number of channels removed.
- 20% (a total of 900,014) of all videos removed were removed for violent or graphic content during this same period.

YouTube’s awareness raising efforts are outlined in response to Outcomes 2 and 3 above.

Outcome 6: Provide safeguards to reduce the risk of harm arising from online **misinformation**

Our Community Guidelines do not allow misleading or deceptive content that poses a serious risk of egregious harm. More information on how YouTube addresses misinformation can be found here and here.

- **YouTube misinformation policies:** Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes certain types of misinformation that can cause real-world harm, like promoting harmful remedies or treatments, certain types of technically manipulated content, or content interfering with democratic processes. From April to June 2022, YouTube removed 122,660 videos (or 2.7% of all videos removed) for misinformation.
- **YouTube election misinformation policies:** Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes misinformation that can cause real-world harm, like certain types of technically manipulated content, and content interfering with free and fair democratic election processes.
- **YouTube COVID-19 medical misinformation policy:** YouTube does not allow content that spreads medical misinformation that contradicts local health authorities' (LHA) or the World Health Organization's (WHO) medical information about COVID-19. This is limited to content that contradicts guidance on treatment, prevention, diagnosis, transmission, and the existence of COVID-19. Note that YouTube's policies on COVID-19 are subject to change in response to changes to global or local health authorities' guidance on the virus. There may be a delay between new LHA/WHO guidance and policy updates given the frequency with which this guidance changes, and our policies may not cover all LHA/WHO guidance related to COVID-19.
- **YouTube vaccine misinformation policy:** YouTube does not allow content that poses a serious risk of egregious harm by spreading medical misinformation about currently administered vaccines that are approved and confirmed to be safe and effective by LHA and the WHO. This is limited to content that contradicts LHA or WHO guidance on vaccine safety, efficacy, and ingredients.

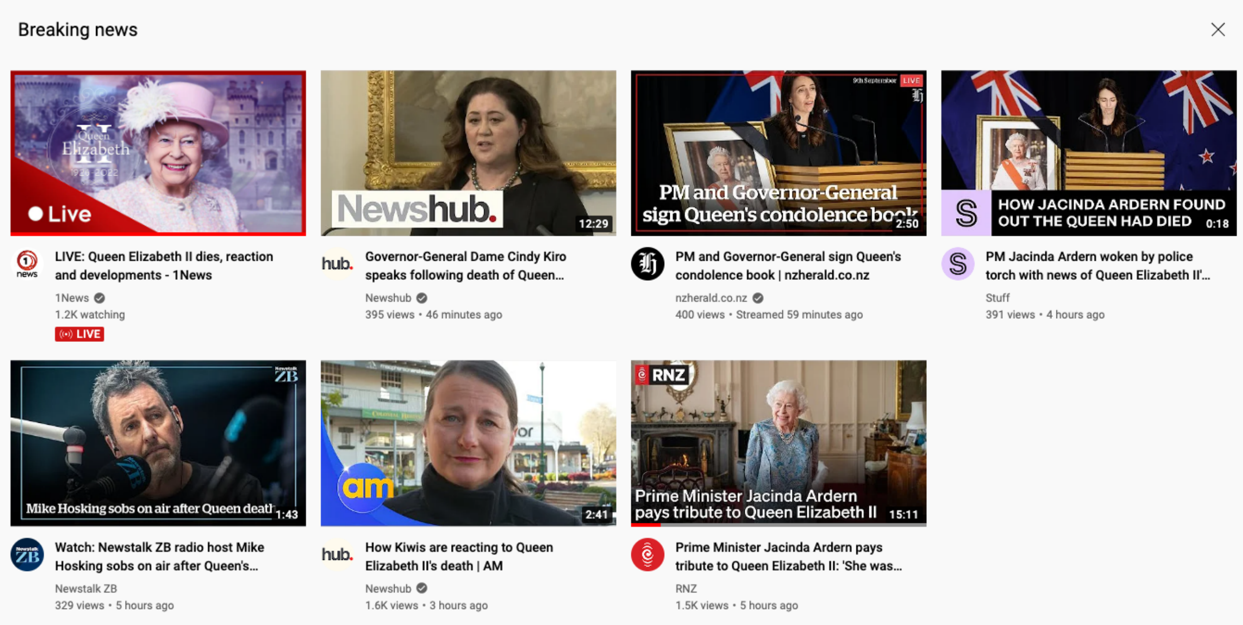
Information on how YouTube enforces our Community Guidelines is set out in the Overview. In addition to our three-strikes system, YouTube has taken steps to strengthen the requirements for monetisation so that spammers, impersonators and other bad actors can't hurt the ecosystem or take advantage of good creators producing high-quality content. Information on the YouTube Partner Program (YPP) is set out in the Overview.

- **YouTube channel monetisation policy**: YouTube monetisation policy includes YouTube's Community Guidelines, Terms of Service, Copyright, and Google AdSense program policies. YouTube enforces this monetisation policy by:
    - Prohibiting or limiting ads against individual videos;
    - Suspending participation in the YouTube Partner Program; and
    - Suspending or terminating a YouTube channel.

In addition to our Community Guidelines, YouTube's recommendation system aims to reduce recommendations of borderline content—i.e. material that comes close to, but does not cross the line of violating our policies. This includes content that may misinform users in harmful ways, such as videos promoting a fake miracle cure for an illness, or videos claiming that the earth is flat. We continue to refine our systems where we operate to address the evolving nature of misinformation. In 2021, YouTube published a blog post, 'On YouTube's Recommendation System,' which provides a detailed explanation of how our recommendation system works, including how we identify harmful misinformation or borderline content.

While YouTube tackles misinformation on the platform by applying the four Rs principles, we also want to support users in thinking critically about the content that they see on YouTube and the online world so that they can make their own informed decisions. We do this in several ways:

- We help users build media literacy skills, including through features and interventions to provide more context to users and ensure that authoritative sources are elevated in response to searches or browsing intents related to health, civic participation, current events or other topics where users want content they can trust.
    - 'Breaking News' on the YouTube homepage and 'Top News Shelves' on YouTube search, prominently surface news content from authoritative sources. For instance, in early 2020 we launched the COVID-19 News shelf on the home page to help connect users with fresh and authoritative content on COVID-19.
    - Authoritativeness in YouTube recommendations prioritise information from authoritative sources for newsworthy events and topics prone to misinformation in search results and recommendations (additional information can be found here). For example:


Breaking news
Queen Elizabeth II 1926-2022
Live
Newshub.
12:29
PM and Governor-General sign Queen's condolence book
2:50
HOW JACINDA ARDERN FOUND OUT THE QUEEN HAD DIED
0:18
ZB
Mike Hosking sobs on air after Queen death
1:43
am
2:41
RNZ
Prime Minister Jacinda Ardern pays tribute to Queen Elizabeth II
15:11

LIVE: Queen Elizabeth II dies, reaction and developments - 1News
1News
1.2K watching
LIVE

Governor-General Dame Cindy Kiro speaks following death of Queen...
Newshub
395 views • 46 minutes ago

PM and Governor-General sign Queen's condolence book | nzherald.co.nz
nzherald.co.nz
400 views • Streamed 59 minutes ago

PM Jacinda Ardern woken by police torch with news of Queen Elizabeth II'...
Stuff
391 views • 4 hours ago

Watch: Newstalk ZB radio host Mike Hosking sobs on air after Queen's...
Newstalk ZB
329 views • 5 hours ago

How Kiwis are reacting to Queen Elizabeth II's death | AM
Newshub
1.6K views • 3 hours ago

Prime Minister Jacinda Ardern pays tribute to Queen Elizabeth II: 'She was...
RNZ
1.5K views • 5 hours ago

- We enable the work of organisations who work on media literacy initiatives. For instance, in 2020, the Alannah and Madeline Foundation launched the Media Literacy Lab thanks to a $1.4m grant from Google. The Media Literacy Lab seeks to empower young people to think critically, create responsibly, be effective voices and active citizens online.

- We invest in thought leadership to understand the broader context of misinformation. As the nature of misinformation rapidly evolves, it is critical that people understand the broader context of misinformation on the internet. Jigsaw, a unit within Google, has developed research, technology, and thought leadership in collaboration with academics and journalists to explore how misinformation campaigns work and spread in today's open societies.

Among a wide range of fact-checking programmes across Asia-Pacific and beyond, Google has also supported Full Fact, a nonprofit that provides tools and resources to fact checkers. Through our philanthropic arm, Google.org, we provided Full Fact with $2 million and seven Googlers from the Google.org Fellowship, a pro-bono program that matches teams of Googlers with nonprofits for up to six months to work full-time on technical projects. The Fellows helped Full Fact build AI tools to help fact checkers detect claims made by key politicians, then group them by topic and match them with similar claims from across press, social networks and even radio using speech to text technology. Over the past year, Full Fact boosted the amount of claims they could process by 1000x, detecting and clustering over 100,000 claims per day — more than 36.5 million total claims per year. The AI-powered tools empower fact checkers to be more efficient, so that they can spend more time actually checking and debunking facts rather than identifying which facts to check. Using a machine learning BERT-based model, the technology now works across four languages (English, French, Portuguese and Spanish). And Full Fact's work has expanded to South Africa, Nigeria, Kenya with their partner Africa Check and Argentina with Chequeado. In total in 2020, Full Fact's fact checks appeared 237 million times across the internet.

At a local level, under the Google News Initiative $3M Open Fund, Google provided funding to New Zealand's Stuff, to support their campaign "The Whole Truth: COVID-19 Vaccination". In partnership with Māori Television and the Pacific Media Network, this project is critical to reaching Māori and Pacific communities in Aotearoa, with accurate and easy to understand information on the vaccine and addressing vaccine misinformation.

We are also working with industry and other stakeholders to respond to evolving harms arising from misinformation:

- Supported Aspen NZ Seminars on Misinformation.
- Sponsored Trusted Media Summit (APAC), where news partners build a community to understand the trends in misinformation and audience behaviour.
- Sponsored an online tracking tool for journalists to understand, monitor and report on ahead of NZ's local elections.
- Partnered with Squiz Kids' media literacy module, "Newshounds" to launch its plug and play resources for teachers, children and their parents in NZ.
- Partnered with Te Rito to fund and provide a training camp for 30 Māori cadet journalists with sessions on verification (focused on fake images and information).
- Digital Skills Training for New Zealand Journalists, using Google's strengths and insight (Pinpoint) to train journalists in how to find and present stories to engage audiences.

Outcome 7: Provide safeguards to reduce the risk of harm arising from online **disinformation**

Our Community Guidelines include tough policies against users that deliberately try to deceive or mislead people.

- **YouTube Impersonation policy**: This policy states that 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.
- **YouTube fake engagement policy**: YouTube does not allow anything that artificially increases the number of views, likes, comments, or other metrics either by using automatic systems or serving up videos to unsuspecting viewers. Content and channels that do not follow this policy may be terminated and removed from YouTube.
- **YouTube spam, deceptive practices, and scam policies**: YouTube does not allow spam, scams, or other deceptive practices that take advantage of the YouTube community. We also do not allow content where the main purpose is to trick users into leaving YouTube for another site. 89% of all channels removed between April 2022 and June 2022 were a result of accounts being dedicated to spam. See our YouTube Community Guidelines Transparency Report for further data.

Information on how YouTube enforces our Community Guidelines is set out in the Overview and steps taken to strengthen the requirements for monetization is outlined under Outcome 6 above.

Besides Community Guidelines, we also have a Threat Analysis Group which tracks actors involved in disinformation campaigns. As set out in the Overview, the actions taken against coordinated influence operation campaigns on our platforms are disclosed in the Quarterly Bulletin (published on our Threat Analysis Group blog).

YouTube includes verification badges (✓ or ✔) on channels where it has been verified as the official channel of a creator, artist, company or public figure. Verified channels help users to distinguish official channels from other channels with similar names on YouTube. This information is provided to YouTube Creators on How To Verify Your YouTube Account.

Google supports responsible political advertising, and expects all political ads and destinations to comply with local legal requirements. This includes campaign and election laws and mandated election 'silence periods' for any geographic areas that they target. Google's broader ads policies apply to all ads, including election ads. Election ads in New Zealand are ads that feature:

- a political party, current elected officeholder or candidate for the New Zealand Parliament; or
- a referendum option up for vote, a referendum option proponent or a call-to-vote once a national referendum is officially declared by an Act or Order in Council.

Specific measures relating to New Zealand election ads include:

- A requirement that all prospective advertisers who wish to run election ads in New Zealand complete the Google verification process.

  - Once Google verifies the advertiser's eligibility to run election ads, they receive an email and an in-account notification. Verifying their identity may require two steps and each step can take up to 5 business days. Our teams are trained to handle this process at scale across and are equipped to respond to related questions from the political parties and candidates participating in, and institutions responsible for, elections.

- Advertisers must follow New Zealand law and applicable Electoral Commission guidance related to disclaimers, including having a clear promoter statement in their ads where required. If required by law, advertisers must obtain authorisation from a political party or candidate before purchasing ads.
- Advertisers must comply with New Zealand law on silence periods.

Our election ads transparency tools provide greater protections to internet users and add important accountability on online advertisers.

YouTube's collaboration with industry and other stakeholders to respond to harms arising from disinformation is outlined in Outcome 6, above.

4.2 Empower users to have more control and make informed choices

Signatories recognise that users have different needs, tolerances, and sensitivities that inform their experiences and interactions online. Content or behavior that may be appropriate for some will not be appropriate for others, and a single baseline may not adequately satisfy or protect all users. Signatories will therefore empower users to have control and to make informed choices over the content they see and/or their experiences and interactions online. Signatories will also provide tools, programs, resources and/or services that will help users stay safe online.

Outcome 8. Users are empowered to **make informed decisions** about the content they see on the platform

Our goal is to connect users with content that they love, so we offer tools to help shape recommendations based on interests. We also provide ways for users to tell us when we're recommending something that they aren't interested in. YouTube provides users with a variety of opportunities to make informed choices about the content encountered. The following are examples of some tools and features we have created:

- For families, YouTube offers options to decide which YouTube experience is best for them, recognising that every family has a different approach to how they use technology, explore online and set digital ground rules. An overview of YouTube Kids, our separate app made just for kids, and our YouTube Supervised Experience, which allows parents to manage their kids' experience on YouTube, is provided in response to Outcome 1. Further information can be found here.
- Verification badges to inform users where we have verified a channel as the official channel of a creator, artist, company or public figure and help users distinguish official channels from other channels with a similar name on YouTube.

- YouTube has recently introduced handles, a new product feature for people to find and engage with creators and each other on YouTube. Handles are unique to individual channels and will make it easier for users to identify specific YouTube channels in comments, community posts, video descriptions and more.

- Restricted mode is an optional setting that you can use to help screen out potentially mature content that you or others using your devices may prefer not to view. Restricted mode was created to give viewers better control over the content that they see. This mode intentionally limits their YouTube experience. Viewers who turn on Restricted mode can't see comments on videos. See here for further detail.

- YouTube allows users to manage their watch history by clearing or pausing their history and by toggling it on and off at any time. Any videos watched while history is paused won't be used to improve a user's video recommendations. Signed in users are also able to view their watch history. Users are also able to delete their search history.

- Similarly, users signed in to a YouTube account are able to remove recommended content (videos, channels, sections and playlists) from their Home. Removing these videos can help improve recommendations.

Our awareness raising and educational efforts to counter misinformation, including YouTube Community Guidelines Enforcement report and the Google News Initiative $3M Open Fund are detailed under Outcome 6.

Outcome 9. Users are **empowered with control** over the content they see and/or their experiences and interactions online

As detailed in our response to Outcome 8, YouTube offers tools to enable users to control the content they see on our platform through supervised accounts for children, using restricted mode and by managing their watch history. In addition to these tools, a user can also subscribe to channels they like to see more content from those channels. Once a user subscribes to a channel, any new videos it publishes will show up in the Subscriptions feed.

YouTube also enables users to to turn personalized ads on or off. If they're turned on, we also allow you to turn on or off particular categories of ads (like apparel, banking, etc). To better protect children, personalized ads, remarketing, and other personalized targeting features are prohibited on YouTube for:

- Google Accounts managed by Family Link for children under the age of 13 (supervised accounts); and

- Content set as made for kids.

Contextual ads can be served on YouTube for supervised accounts and on content set as made for kids. These ads are based on factors like:

- The content being viewed.

- The viewer's current search.

- The viewer's general location (such as city or state).

Ads must follow the made for kids ad policy to be eligible to appear on YouTube for supervised accounts and on content set as made for kids.

Viewers of "made for kids" content may see an ad bumper before and after a video ad is shown. This bumper helps alert them when an advertisement is starting and ending. If viewers have a YouTube Premium family plan, their children are eligible for ad-free content and other shared benefits of membership.

4.3 Enhance transparency of policies, processes and systems

Transparency helps build trust and facilitates accountability. Signatories will provide transparency of their policies, processes and systems for online safety and content moderation and their effectiveness to mitigate risks to users. Signatories, however, recognise that there is a need to balance public transparency of measures taken under the Code with risks that may outweigh the benefit of transparency, such as protecting people’s privacy, protecting trade secrets and not providing threat actors with information that may expose how they may circumvent or bypass enforcement protocols or systems.

Outcome 10. Transparency of policies, systems, processes and programs that aim to reduce the risk of online harms

The success of our business is built on providing trusted products and services, and transparency about how we organize content is essential to that trust.

Our policies work best when users are aware of the rules and understand how we enforce them. That is why we work to make this information clear and easily available to all. We provide a comprehensive help center with detailed information about policies—including our Community Guidelines and Advertising-Friendly Guidelines—along with blog posts that detail the specific provisions of our policies. In addition, we regularly release reports that detail how we enforce those policies or review content reported to be in violation of local law.

As part of our ongoing commitment to transparency, we’ve created How YouTube Works — a website designed to answer the questions we most often receive about what we're doing to foster a responsible platform for our community, and explain our products and policies in detail. The website addresses some of the important questions we face every day about our platform — involving topics such as child safety, harmful content, misinformation, and copyright, as well as tackling timely issues as they arise, like how we have responded to the COVID-19 crisis and how we support elections. Within the site, we explain how we apply our responsibility principles (the four Rs) — which work alongside our commitment to users’ security — to tackle these important questions. How YouTube Works provides an in-depth look at our products and settings, such as YouTube Search, Recommendations, privacy controls, and Ad Settings, showing how they help people have the best possible experience while they’re using YouTube. Additionally, users will find details of our policies — like our Community Guidelines and monetization policies — so everyone in the community knows what they can and can’t do on YouTube. We explain how our policies are developed and enforced in partnership with a wide range of external experts and Creators.

The YouTube Official Blog also provides further detail on how YouTube works, including the development and evolution of our policies and products.

Outcome 11. Publication of regular **transparency reports** on efforts to reduce the spread and prevalence of harmful content and related KPIs/metrics

Since Google launched its first Transparency Report in 2010, we’ve been sharing data that sheds light on how the policies and actions of governments and corporations affect privacy, security, and access to information online.

The YouTube Community Guidelines Enforcement Transparency Report provides quarterly updates on the number of videos, channels, and comments removed from YouTube, including a breakdown of the policies under which this content was removed. It also details how we detect infringing videos (e.g., with automated systems, via user flags) and how many offending videos were removed without any user viewing them.

A list of our transparency reports (including links to the materials) is provided in the Overview. Our website How YouTube Works (see Outcome 10 above) provides detailed information about how we use algorithms to rank and recommend content.

Google also makes available data on Government requests to remove content, for instance where a Government body claims that content violates local law or a government requests that we review content to determine if it violates our own product community guidelines and content policies. Data on removal requests for New Zealand can be found here.

4.4 Support independent research and evaluation

Independent local, regional or global research by academics and other experts to understand the impact of safety interventions and harmful content on society, as well as research on new content moderation and other technologies that may enhance safety and reduce harmful content online, are important for continuous improvement of safeguarding the digital ecosystem. Signatories will seek to support or participate in these research efforts.

Signatories may also seek to support independent evaluation of the systems, policies and processes they have implemented under the commitments of the Code. This may include broader initiatives undertaken at the regional or global level, such as independent evaluations of Signatories’ systems.

Outcome 12. Independent research to understand the impact of safety interventions and harmful content on society and/or research on new technologies to enhance safety or reduce harmful content online.

YouTube continues to support global and local research efforts, such as our partnership with Squiz Kids’ media literacy module; “Newshounds“ to launch its plug and play resources for teachers, children and their parents in NZ; and with Te Rito to fund and provide a training camp for 30 Māori cadet journalists with sessions on verification (focused on fake images and information). In addition, YouTube is a supporter of NetSafe’s programs and initiatives to reduce online harm.

Google also is a member of the Digital Trust & Safety Partnership, a first-of-its-kind partnership with other leading technology companies committed to developing industry best practices, verified through internal and independent third-party assessments, to ensure consumer trust and safety when using digital services. As discussions on these important issues continue, the group will engage with consumer and user advocates, policymakers, law enforcement, relevant NGOs and various industry-wide experts to help us develop these best practices. The Partnership will share a state-of-the-industry report that evaluates companies' implementation of the practices.

In addition to our publicly available transparency reports (see Outcome 11), YouTube is equipping researchers from around the world with data, tools, and support to advance the public's understanding of our platform and its impact through our recently launched YouTube Researcher Program. Eligible researchers from diverse disciplines can apply to use YouTube data to study a variety of topics. We're starting this program offering participants the following:

- Scaled access to YouTube's public data corpus with as much quota as required for their research.
- Opportunity to derive insights from global YouTube data.
- Support and technical guidance from YouTube.

We will continue to sponsor the joint annual conference between NZ's Netsafe and the AU's eSafety Commissioner.

Outcome 13. Support independent evaluation of the systems, policies and processes that have been implemented in relation to the Code.

As per Measure 45 of the Code, Google is committed to selecting an independent third-party organization to review its annual compliance reports under the Code and evaluate the level of progress made against the Commitments, Outcomes and Measures, as outlined in section 4 of the Code, as well as commitments made by Signatories in their Participation Form.