Information quality & content moderation

Google YouTube

Introduction

The internet has been a powerful force in its relatively brief existence. Engineered to shuttle information from one computer to the next, without regard for its contents or intent, it quickly became the backbone of the modern era. Today, it connects individuals and communities around the world. It can inspire the best of society by democratizing access to knowledge, powering business, and providing new opportunities for art and creativity.

At Google in 2020, as we progressed on our vision to build a more helpful Google for everyone, this has meant hundreds of billions of search queries, 500 hours of content uploaded to YouTube every minute, billions of monthly direct connections between businesses and customers, and many hundreds of billions of dollars of economic activity.[1] Behind these numbers are countless stories of discovery and connection, communities finding support, and people finding answers to their questions big and small.

We feel a great responsibility to our users when they place their trust in us to deliver them trustworthy, helpful information that meets their needs. Like every form of communication before it, the internet can be misused. The same low barriers to entry that fueled its growth around the world have enabled its manipulation by bad actors seeking to inflict harm on others, whether seeking profit or promoting ideology. The right way to manage a decentralized internet to favor the good while reducing the bad has been actively debated since its creation.

One focus of these discussions is the content people share – text, images, videos, and web pages – generated in increasing frequency and made accessible ever more rapidly around the world.

Our mission at Google is to organize this information and make it universally accessible and useful. Core to this mission is a focus on the relevance and quality of the information we present to users. In different ways across our different platforms, we strive to connect people with 'high-quality information'; the most useful, trustworthy, and helpful content at the moment a person needs it. At the same time, we work to prevent user and societal harm and limit the reach of 'low-quality information'; content that strays furthest from those qualities.

[1] See e.g., https://economicimpact.google.com/; https://www.blog.google/around-the-globe/google-europe/helping-europeans-succeed-googles-impact-in-europe/

Sorting ‘high-quality’ from ‘low-quality’ information is a large, dynamic challenge without a perfect answer. The breadth of information available online makes it impossible to give each piece of content an equal amount of attention, human review, and deliberation. Even if that were possible, reasonable people could disagree on appropriate outcomes. Similarly, no ranking system can be perfect, nor will everyone agree on the values for which they should optimize.

Still, this is a challenge we are dedicated to tackling–as we have been since Google’s founding. With this paper, we aim to share our approach to information quality and content moderation. It outlines the key considerations that guide our product, policy, and enforcement decisions, as well as the four complementary levers we use to implement those principles across our services. It also provides a look into the vital work being done in collaboration with other technology companies, civil society, academia, and government to support information quality across the internet.

We welcome and look forward to the feedback we will receive in response to this paper and will continue to provide additional transparency on this topic in the future.

Our approach to information quality

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.

The product, policy, and enforcement decisions we make in this complex environment are guided by a set of considerations that are consistent across the spectrum of our products and services:

- **Value openness and accessibility:** We aim to provide access to an open and diverse information ecosystem. But that doesn't mean that anything goes on our services. As we will describe later in the paper, removal of content is an important lever we use to address information quality. However, it is not the only lever at our disposal, and we use it with caution, particularly in the context of Search. We believe that a healthy and responsible approach to supporting information quality should aim toward keeping content accessible.

- **Respect user choice**: Users who express an intent to explore content that is not illegal or prohibited by our policies should be able to find it, even if all available indicators suggest it is of relatively low quality. We set a higher bar for information quality where users have not clearly expressed what they are looking for.

- **Build for everyone**: Our services are used around the world by users from different cultures, languages, and backgrounds, and at different stages in their lives. Some have always known a world with smartphones, while others have lived most of their lives without access to the web. Our product and policy development, as well as our policy enforcement decisions, take into account the diversity of our users and seek to address their needs appropriately.

These have been priorities since our founding. They have guided our evolving approach toward information quality, taking into account shifting user expectations and norms, increasing sophistication of malicious actors, and the evolving nature of the web.

Each of the products and services we offer has a different purpose, and people have different expectations of what kind of content they will interact

with on each. So, we tailor our approach to the content that should be available on each product and service carefully.

Our products and services fall on a spectrum, from least to most restrictive. Google Chrome is a tool for viewing the breadth of content on the internet, warning only of pages potentially infected by malware. Google Search serves as an index of all pages available on the open web, where users expect to find every legal webpage pertaining to their query. Therefore, it leans toward the least restrictive end of that spectrum. On the other end, our advertising products are among the most restrictive, as we do not want to profit from those who create harmful content or experiences. Other products fall elsewhere on the spectrum. For instance, Gmail involves minimal limitations on content, while YouTube is a platform for uploading and sharing content as part of a community, which requires broader prohibitions than Google Search.

Similarly, specific features within a product may fall at different points on that spectrum. For instance, some features of Google Search, like Autocomplete, provide information to help people get to the results they are looking for as quickly as possible.[2] But we also want to be careful not to show potentially upsetting content to people when they haven't asked for it. For these features, we have developed policies to exclude things like pornography, hate speech, or violence from appearing. Actions taken on these features do not limit what users can search for.

We rely on four complementary levers to support information quality and moderate content across many of our products and services:

- **Remove**: We set responsible rules for each of our products and services and take action against content and behaviors that infringe on them. We also comply with legal obligations requiring the removal of content.
- **Raise**: We elevate high-quality content and authoritative sources where it matters most.
- **Reduce**: We reduce the spread of potentially harmful information where we feature or recommend content.
- **Reward**: We set a high standard of quality and reliability for publishers and content creators who would like to monetize or advertise their content.

These levers allow us to be consistent in our methodology across products while tailoring their implementation to fit the uses and needs of each. In the following sections, we explore how each of them works in practice.

[2] https://support.google.com/websearch/answer/106230?hl=en

As such, they have guided our response to the Coronavirus pandemic in 2020 – a summary of our response to Coronavirus misinformation is available immediately before the conclusion of this paper.

Remove: developing and enforcing our ‘rules of the road’

One lever we deploy in our effort to support information quality on our products and services is the removal of content from a given platform entirely. Removal of content may occur for two reasons: it violates the law, or it violates the ‘rules of the road’ for that product or service.

We comply with the law in each country in which we operate and remove illegal content on our platforms in that country. In every country in which we operate, the unique cultures, histories, and forms of government have produced different laws governing what is considered permissible expression. For instance, in France, Austria, and Germany, regulatory frameworks prohibit denial of the Holocaust. Some countries provide individuals with broad rights against alleged defamation, while others take a more limited view. The European Union and Russia have adopted data protection regimes that afford individuals a so-called “right to be forgotten” by requesting platforms to delist specific outdated material about them.

In addition, we develop and maintain ‘rules of the road,’ which outline what types of content and behaviors are acceptable for each product or service. Known as ‘content policies’ or ‘community guidelines,’ we aim to make them clear and easily accessible to all users and content creators – whether those are video creators, webmasters, app developers, or advertisers. These ‘rules of the road’ articulate the purpose and intended use of a given product or service and represent a crucial part of what makes that product unique. They also explain what types of content and behaviors are not allowed, and the process by which a piece of content, or its creator, may be removed from the service.

Let’s take a look at how these policies are developed and enforced.

Content removals at scale

We enforce our content policies at scale and take tens of millions of actions every day against content that does not abide by the ‘rules of the road’ for one or more of our products.

- In 2019, **more than 30 million videos** were removed from YouTube for violating our community guidelines.
- In 2019, we removed more than **2.7 billion bad ads** from our systems and took action against almost **1 million** bad advertiser accounts. On the publisher side, we terminated over 1.2 million accounts and removed ads from over 21 million web pages that are part of our publisher network for violating our policies.
- Google Play's policies prohibit numerous types of deceptive behaviors and misleading content, especially when it relates to the dissemination of applications related to medicine or personal health. When developers are found to infringe on these policies, their apps may be removed from the Google Play store. Throughout 2019, Google Play stopped over **790,000** policy-violating apps before they were ever published to the Play Store.
- In 2019, Google Maps detected and removed more than **75 million** policy-violating reviews and **4 million** fake business profiles, and took down more than **580,000** reviews and **258,000** business listings that were directly reported to us for violating our policies. We also reviewed and removed more than **10 million** photos and **3 million** videos that violated our content policies on Google Maps, and disabled more than **475,000** user accounts that were found to be abusive.

Developing policies and designing for safety

We design the 'rules of the road' across all our products and services to protect users from harm while supporting the purpose of the product. For each product and service, we tailor these policies to strike the appropriate balance between providing access to a diversity of voices and limiting harmful content and behaviors.

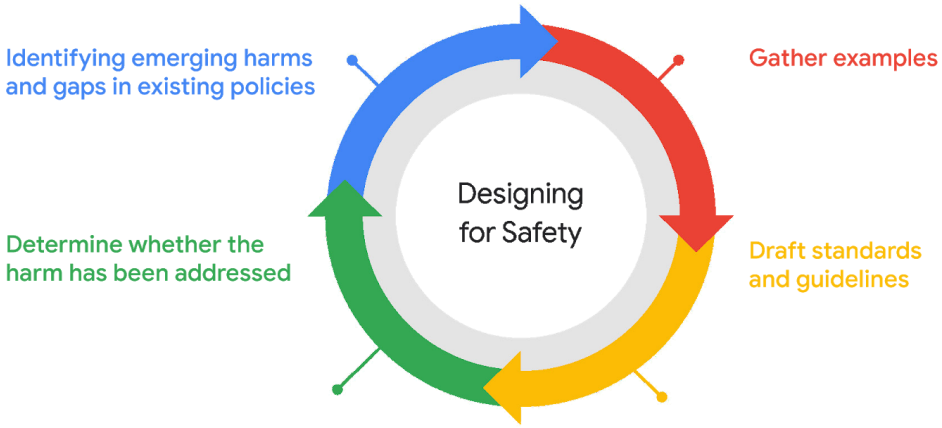
This balance can differ from one product to the next, in part because harm manifests differently in each service and context. While a universally recognized harm may be prohibited across all our products and services, it can appear on each product and service differently. So, we must evaluate the potential for harm specific to each product and design our policies accordingly. This includes harm to an individual and harm that may affect an entire society, such as an attempt to interfere with elections or civic processes.

Among others, we consider the following types of risks when considering what safeguards and rules may be needed for each product and service:

- **Encouraging harmful or dangerous behavior**: content that either depicts particularly harmful or dangerous behaviors, or encourages users to engage in those behaviors.
- **Hateful content**: Content that promotes or condones violence against individuals or groups based on characteristics like race, ethnicity, gender identity, religion, and veteran status.
- **Threats, harassment, and bullying:** Content that involves direct threats to others, blackmail, exposure of private data, or is intended to harass or silence.
- **Violent or graphic content**: Content for which the primary purpose is to be shocking, sensational, gratuitous, or offensive, including content produced by, or in support of, a terrorist organization.
- **Sexually explicit content:** Written or visual depictions of nudity or graphic sex acts, with the exception of nudity for educational, documentary, or scientific purposes.
- **Spam, abuse, and deceptive practices**: Activities that attempt to abuse our products, circumvent protections to safeguard user data, manipulate ranking systems, or cause broadly invalid traffic that doesn't derive from genuine user interest.
- **Impersonation, misrepresentation, and scams**: Activities that misrepresent an individual's identity, place of business, country of operations, or the sale of goods and services.

To help us identify emerging harms and gaps in our existing policies, we consider expert input, user feedback, and regulatory guidance. We rely on research performed by analysts who study the evolving tactics deployed by bad actors, trends observed on other platforms, and emerging cultural issues that require further observation. We also engage in conversations with regulators around the world. Their perspectives and concerns directly inform our policy process.

Policy creation process



Next, we gather as many examples of how a particular harm has manifested on our services, or might manifest in the future, and look for common threads. We also consider counter-examples of content that may look similar to the harmful content we wish to address, but is actually benign or of significant public interest. This helps us define the common traits that make the content or behavior harmful, as well as the risks that an overbroad policy would pose.

With that, we develop draft standards and enforcement guidelines, test them against the counter-examples to minimize false positive enforcement, consult with many experts across disciplines at Google, and further consider perspectives from experts outside of Google. We then work to resolve conflicts within the diverse feedback and synthesize the draft standards and guidelines into coherent policy. Finally, we 'incubate' policies by testing them until we are confident that we can ensure a high level of consistency in their application before rolling them out further.

We continue this process of exploration and refinement until we have an approach that is clear, predictable, and repeatable. We strive to ensure that reasonable users or content creators, upon being informed of any change to our policies, can understand what it refers to (clear), and can determine whether their content or behavior is likely to be affected by this rule (predictable). The new rule should also be sufficiently generic that it can be applied consistently across multiple independent cases globally (repeatable).

Finally, before we begin implementation and enforcement of the new policy, we determine whether it has addressed the harm it targeted, measure the impact of the change on existing users, assess how to provide proper notice of the change, and provide the proper mechanisms for enforcement.

This is a time-consuming process. It can take months before we feel comfortable taking action on a new policy. This collaborative approach taps into multiple areas of expertise within and beyond our company and is typically

driven by our Trust and Safety teams. Their mission includes tackling online abuse by developing and enforcing the policies that keep our products safe and reliable. The team includes product specialists, engineers, lawyers, data scientists, and others who work together around the world and with a network of in-house and external safety and subject matter experts.

As we engage in this process, we know that some may disagree with the decisions we have made in our attempt to strike the right balance between reducing harm and upholding principles of user access and choice. There is rarely a simple, correct answer to these questions, and reasonable people can disagree on how to moderate content. This is especially true as the internet ecosystem evolves and users' expectations change with them. We believe that an inclusive process, transparency in our work, and a willingness to reassess our policies will best serve our users and the societies in which we operate.

Developing a policy to prohibit speculative and experimental medical treatments in Google Ads

While developing our policies, we weigh multiple viewpoints and considerations. This was true, for instance, in September 2019 when we updated our Ads policies to prohibit speculative and experimental medical treatments, including stem cell therapy or gene therapy.[3]

While we are aware that important medical discoveries often start as unproven ideas, we must also consider the impact to the safety of our users. Although several treatments had been approved by some regulatory bodies, we observed a rise in bad actors attempting to take advantage of individuals by offering untested and deceptive treatments. The deception could cause people to spend large amounts of money on unproven treatments that may not provide a medical benefit, or could even cause serious health issues. After consulting third-party experts, we moved to prohibit ads for these treatments to prevent our advertising platforms from being misused in ways that could lead to serious financial and physical harm to our users.

Proactive detection & enforcement

By the numbers

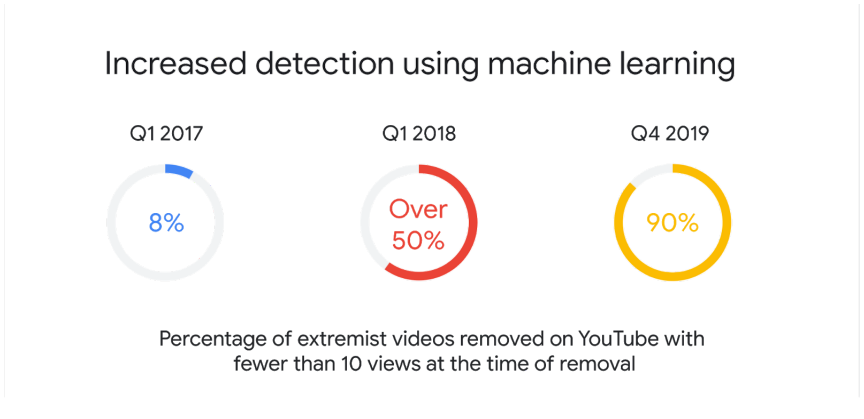
From January to March 2020:

[3] https://support.google.com/google-ads/answer/9475042?hl=en

- **More than 6.1 million videos** were removed from YouTube for violating our community guidelines.
- **93%** of these videos were first flagged by machines rather than humans. Of those detected by machines, **53%** never received a single view, and just over **81%** received fewer than 10.
- In this same period, YouTube removed more than **693 million comments,** the majority of which were spam. 99% of removed comments were detected automatically.

Case study: using machine learning to help detect extremist content on YouTube

- We introduced machine learning technology to detect extremist content on YouTube in June 2017. To train our machine learning classifiers, our teams reviewed **over 2 million pieces of content.**
- **In Q1 2017, 8%** of videos removed for violating the violent extremism policy had fewer than 10 views at the time of removal.
- By **Q1 2018,** that figure reached more than **50%.**
- In **Q4 2019**, approximately **90%** of the videos uploaded that were removed for violating the violent extremism policy were taken down before they had 10 views.



To enforce our policies at the scale of the web, we rely on a mix of automated and human efforts to spot problematic content. In addition to flags by

individual users, sophisticated automated technology helps us detect problematic content at scale. Our automated systems are carefully trained to quickly identify and take action against spam and violative content. This includes flagging potentially problematic content for human reviewers, whose judgement is needed for the many decisions that require a more nuanced determination. The context in which a piece of content is created or shared is an important factor in any assessment about its quality or its purpose. We are attentive to educational, scientific, artistic, or documentary contexts, including journalistic intent, where the content might otherwise violate our policies.

In addition, our expert teams around the world handle the investigations of more sophisticated threat actors that are adept at circumventing the automated defenses we build into our products. New forms of abuse and threats are constantly emerging that require human ingenuity to assess and plan for action before an automated system can address them at scale. So, we operate dedicated threat intelligence and monitoring teams (e.g., Google's Threat Analysis Group[4]) which provide insights and intelligence to our policy development and enforcement teams so they can stay ahead of bad actors.

Over the past two decades, we have invested in and refined our approach to detection and enforcement at scale. However, because of the open nature and scale of our products and the web, and because motivated bad actors are nimble and not often deterred, catching all problematic content and activity with perfect accuracy is not feasible. We could expand our detection efforts by relying more heavily on our automated systems to catch more content faster. But, this comes with trade-offs.

If we expand our detection efforts in this way, we increase the risk of 'false positives,' or incorrectly removing a piece of content that does not actually violate our policies. This might include important expressions from diverse voices, or content of relevance to the public interest. Its removal could then introduce confusion for content creators and other partners who need our policies to be clear.

We also risk diverting the attention of our Trust and Safety enforcement teams to content that may be innocuous, giving more room for savvy bad actors to slip under the radar. Conversely, if we were to focus our efforts on a narrower set of challenges, we would risk missing the bigger picture, creating blind spots that others could exploit while introducing harm to our users.

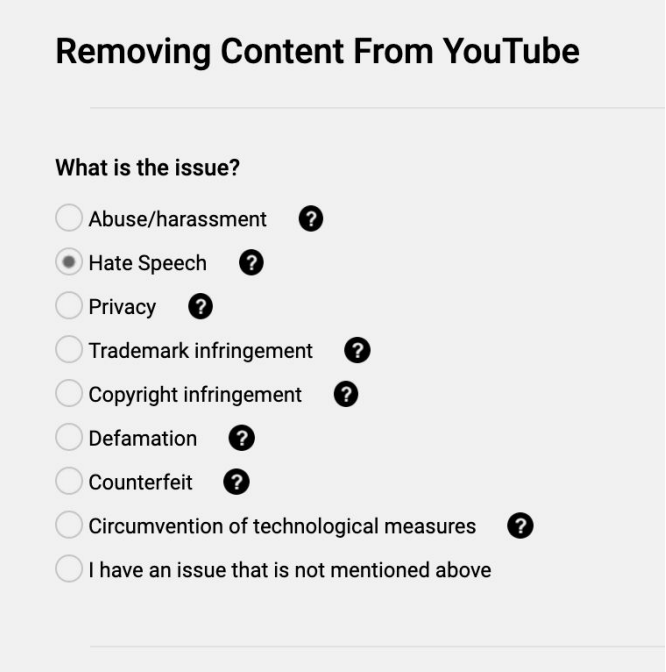
While there is no silver bullet to address this challenge, we are aware of the responsibility that comes with operating at this scale. A wrong decision can have a significant impact on our users, developers, creators, or advertisers. As such, we continue to develop the tools and resources that comprise our detection and enforcement efforts.

[4] https://blog.google/threat-analysis-group

To complement our own efforts, we enable users and trusted organizations to flag content that may be problematic. We take action on content flagged by users after it has been reviewed by a member of our Trust and Safety team to ensure the content does indeed warrant action.

In addition to individual pieces of content and behaviors, we have dedicated responses to take on bad actors themselves. By taking action at the account level when faced with severe or repeated policy violations, we address the root cause of infringements of our policies, and better protect our users. For instance, in our Bad Ads reports, we have described how using machine learning technology allowed us to identify and terminate bad advertiser accounts.[5]

Enabling User Feedback



Sometimes, bad actors try to evade detection of their efforts by using private channels for sharing content. We continue to work with safety and privacy experts to ensure we are using best-in-class techniques to both improve detection and respect the privacy of our users.

Providing transparency into our policy enforcement

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.

[5] For 2020: https://www.blog.google/products/ads/stopping-bad-ads-to-protect-users/

We develop comprehensive help centers, community guidelines websites, and 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.

- 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.[6]

- Our annual **'Bad Ads' report** outlines the scale of our work to enforce our advertising policies, including the number of ads that were removed, the number of pages that we stopped showing ads on, the number of advertiser and publisher accounts that were terminated throughout the year, and the number of updates we made to our policies over the course of the year.[7]

- Our Threat Analysis Group's **Quarterly Coordinated Influence Operations Bulletin**[8] provides information about actions we take against accounts that we attribute to coordinated influence campaigns (foreign and domestic).

- Reports made available on the **Google Transparency Report Website** provide information regarding government requests to remove content from our services, and how the actions of governments and corporations affect privacy, security, and access to information online.[9]

- We also provide a publicly accessible, searchable, and downloadable **Google Transparency Report of election ad content** and spending on our platforms.[10] Given recent concerns and debates about political advertising, and the importance of shared trust in the democratic process, we hope to improve voters' confidence in the political ads they may see on our ad platforms.

We will continue building upon these transparency efforts in the future, as they are an important component of ensuring an informed public dialogue about the role that our services play in society.

Appealing an enforcement action

[6] https://transparencyreport.google.com/youtube-policy/removals
[7] https://www.blog.google/products/ads/enabling-safe-digital-advertising-ecosystem/
[8] See all bulletins on the Threat Analysis Group Blog: https://blog.google/threat-analysis-group
[9] https://transparencyreport.google.com/
[10] https://transparencyreport.google.com/political-ads/home?hl=en

Sometimes, we make mistakes in our decisions to enforce our policies, which may result in the unwarranted removal of content from our services. To address that risk, wherever possible, we make it clear to creators that we have taken action on their content and provide them the opportunity to appeal that decision and give us clarifications. The decision will then be evaluated by a different member of our Trust and Safety team.

Appeals under our misrepresentation policy

We review ads using a combination of automatic and manual processes, and occasionally re-review them for compliance. In some scenarios, an advertiser's business model may, upon initial review, be considered non-compliant with our policies. This can happen when we cannot determine the details of the services or products offered. In these cases, we may verify policy compliance by learning more about the advertiser's business practices as a whole when they appeal our decisions. For example, our Misrepresentation Policy ensures that ads do not deceive our users.[11] When claims of relationships with another business cannot be verified directly from the advertiser's landing page, the ads may initially be disallowed. Additional evidence may be able to support and confirm the validity of such claims. In some advertiser appeals, the advertiser has provided us with additional information that helps us better understand their business model and service delivery chain. This can include official documentation proving their relationship with another business or their official participation in an affiliate network.

We want to make it easy for good-faith actors to understand and abide by our rules, while making it challenging for bad actors to flout them. That is why we seek to make room for good-faith errors as we enforce our rules.
We recognize that anyone can inadvertently take a joke too far or not immediately realize the problematic nature of something they have done or shared. For example, if an individual app infringes on our policies, we typically take action on that specific app rather than sanctioning the account of the developer.

On the other hand, in cases of serious, repeated, or deceptive violations, we may take action that affects an entire website, channel, or app. In the most serious cases, we will shut down user accounts.

[11] https://support.google.com/adspolicy/answer/6020955?hl=en

On **YouTube**, some violations of our community guidelines may result in a 'strike' which restricts a creator's ability to post or create content on the platform for one week. If the creator's behavior warrants another 'strike' within 90 days from the first, a new two-week prohibition from posting or creating content is implemented. A third strike within 90 days results in permanent removal of a channel from YouTube. Creators can appeal those strikes if they believe we are mistaken.

Between January and March 2020, **more than 1.9 million channels** were removed from YouTube for violating our community guidelines.

Supporting reviewer wellness

As we work to reach our goals on information quality and content moderation we rely heavily on machines and technology, but human reviewers play a critical role. These reviewers perform over billions of reviews every year, working to make the right enforcement decisions and helping build training data for machine learning models.

While most content moderation is not violent or graphic, some of the material these moderators review can be disturbing and upsetting. Some moderators chose to work in areas that might be particularly challenging because they seek to have a positive impact on finding and removing this content from the web.

To assist them, we use technology to take on some of the hardest tasks. Today, automated flagging allows us to identify and act more quickly and accurately to remove content, lessening both the burden on human reviewers and the time it takes to remove violative content. For example, more than 90% of the videos we removed from YouTube for violating our community guidelines in Q4 2019 were first flagged by our automated systems. We are constantly making those systems better and more accurate.
The people who review this content do vital work to keep digital platforms safer for everyone, and it can be difficult or emotionally challenging. Google is determined to support the wellness of these workers through a comprehensive wellness program, verification of vendors' compliance with those standards, and research and technological innovation to promote wellness and ensure that those doing this work have access to the resources they need for their wellbeing and mental health.

Wellness standards
Content moderators help us assess context and nuance, to evaluate content we've never seen before, and make distinctions and decisions. We are committed to ensuring they have the highest standard of support and have invested significantly in these teams.
We do this by:

- Providing access to on- and off-site counselling for workers who need it via individual and group sessions, dedicated wellness spaces, and 24/7 phone or on-site counsellor support.
- Limiting work hours for those focusing on sensitive content: reviewers moderating sensitive content also work abbreviated hours, spending no more than 5-6 hours reviewing content in an 8-hour work day.
- Providing the ability for reviewers to opt-out of viewing highly egregious content.
- Providing for physical well-being activities (both available as opt-in and scheduled).
- Providing access to quiet rooms and community spaces, which are required at all sites.

Verification of compliance
We work with third-party vendors and contractors to help us scale our content moderation efforts, and provide the native language expertise and the 24-hour coverage required of a global platform. When we work with these providers, we engage in regular site visits and audits to ensure that our guidelines and Supplier Code of Conduct are respected. Those visits include one-on-one conversations and focus groups with reviewers to provide for direct and confidential feedback. All the third parties we work with provide their employees with grievance reporting and redressal fora, as well as with access to an ombudsperson. We also give employees of our vendors access to the same helpline as Google employees to report concerns, including the option to report anonymously.

Research & technological innovation
In addition to gathering feedback directly from workers and soliciting professional input and advice, we are committed to driving industry-leading research and technological innovation in the field of content moderation. For instance, we published a research paper in 2019 detailing how the use of "grayscale transformations" (converting an image to black and white) can help reduce the emotional impact on moderators. Our research tells us that moderators reviewing violent and extremist content reported an improvement in emotional wellbeing when reviewing content with grayscaling turned on. Given

these findings, we've now built grayscaling into review tools. Because every reviewer is different, grayscaling is an option left open to reviewers, giving them more flexibility when performing reviews. Today, 70% of moderators reviewing violent extremist content on Google Drive, Photos, and others choose to review images in grayscale and keep the grayscale option turned on. We're committed to rolling out this option more broadly.

Grayscaling has its limitations and the same positive effect was not true for all reviewers working on all types of content. That's why we continue to investigate other areas. For instance, blurring content during a review is an approach we thought could be helpful for reviewers. Instead, many reviewers reported feeling nauseous or even irritated from blurring, and preferred to toggle the option off. We're now experimenting with a slider to give moderators the option to adjust the level of blurring when reviewing content and an option that gives them the ability to mouse-over content to unblur key parts of an image. Early results are promising and may lead to a positive impact for some moderators.

More needs to be done to understand the long-term emotional impact of this work. We're conducting new research in 2020 and will continue to share our findings and collaborate closely with the industry.

Content moderation is a relatively new industry and the number of people working in this area has grown significantly in recent years – including within our abuse-fighting teams. This expansion has been an important component of our ongoing work to combat malicious actors and to protect our users from harmful content online. However, we also have an important responsibility to take care of this growing abuse-fighting team as it helps keep our users safe. We are committed to continuing our efforts on both of these fronts.

Raise: Connecting users to authoritative content

Whether on Google Search, YouTube, Google Play, Google Maps, or other consumer services, our products meet user needs by sifting through immense amounts of information. This information comes from well established publishers, new and emerging creators, and individual users who create content as a part of their online journey via comments, reviews, public forums, and social media. We use algorithms to organize that content according to our best understanding of usefulness in addressing the intent and needs of our users.

- The Google Search index represents **more than 100 million gigabytes** of data, mapping hundreds of billions of webpages. If it were to be printed out as books and stacked, there would be enough for 12 round trips to the moon.
- There are billions of Search queries around the world every day, and **15% of the searches** we see each day are searches we've never seen before.
- More than **500 hours of content** are uploaded to YouTube every minute.

To determine whether a piece of content is useful, we must first establish a user's intent. That intent may have been expressed by typing something into a Google Search bar or by watching a cricket match on YouTube, thus passively showing an interest in videos about other cricket games or sporting events.

Usually, multiple pieces of content are relevant to a user's intent, which is why we look to a variety of other factors to rank these pieces of content. Our ranking algorithms look for signals that indicate the expertise, authoritativeness, and trustworthiness of every piece of content so that the best results for the user at that time are at the top. One early and well-known example of this type of algorithm is PageRank, which uses links on the web to assess the importance of a given website.

We are constantly improving these ranking systems. In 2019, Google Search ran more than 383,605 tests to measure the quality of search results and launched more than 3,600 updates to the algorithms that produce them.[12]

[12] For more information on these tests, how they are run, and how we rate Google Search results: see www.google.com/search/howsearchworks

To ensure Search algorithms meet high standards of relevance and quality, we have a rigorous process that involves both live tests and thousands of trained external Search Quality Raters from around the world.[13] Our Search Quality Rater Guidelines which we first published in 2013, define the goals of our ranking systems as they evolve over time, and include the criteria that our raters use to assess the expertise, authority, and trustworthiness of pages.[14] The ratings provided by our Search Quality Raters help us benchmark the quality of our results so that we can meet a high bar for users of Google Search all around the world.

The authority or scientific accuracy of a page is not equally important to all user experiences or all contexts. When a user searches for, or interacts with, entertainment content on YouTube, the reliability of that content matters less. On the other hand, when a user interacts with content related to topics such as their livelihood, civic participation, or news, the trustworthiness of the content provided in response matters considerably more. In such contexts, the health, financial stability, future happiness, or safety of an individual may be directly affected by unreliable information. We refer to these types of topics as "Your Money or Your Life" (YMYL).[15]

For these "YMYL" topics, we assume that users expect us to operate with our strictest standards of trustworthiness and safety. As such, where our algorithms detect that a user's query relates to a "YMYL" topic, we give more weight in our ranking systems to factors like our understanding of the authoritativeness, expertise, or trustworthiness of the pages we present in response. For example, when a user is looking for specific medical information or advice, we work to provide content from authoritative sources like health professionals and medical organizations.

Elevating authoritative information

We develop features or ranking changes that elevate authoritative information, including:

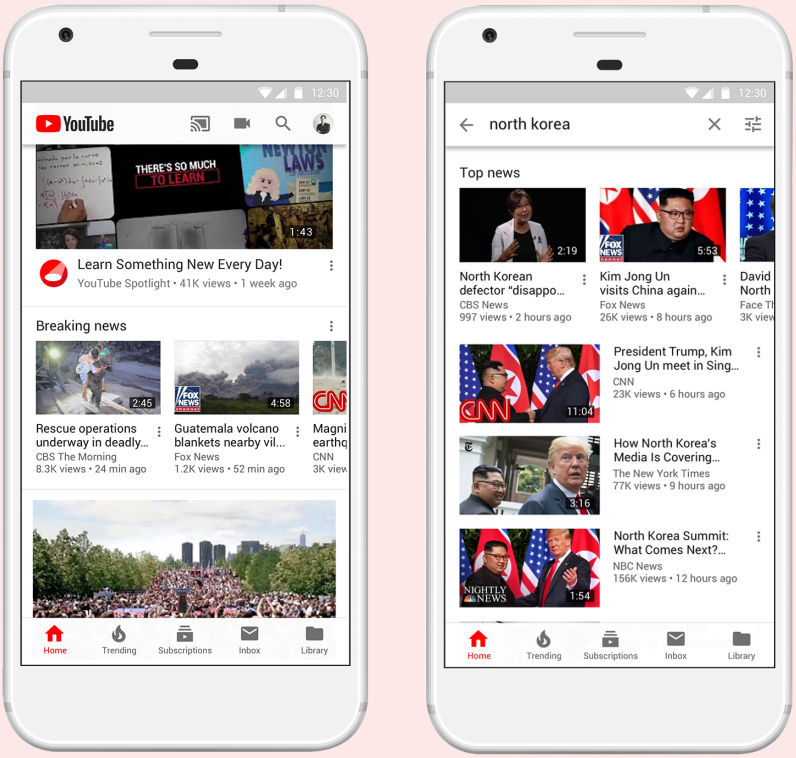
- **TOP AND BREAKING NEWS SHELVES ON YOUTUBE:** On **YouTube**, a **'Top News shelf'** and a **'Breaking News shelf'** prominently display authoritative political news information. The Top News shelf triggers in response to searches that have political news-seeking intent and provides content from verified news channels. The Breaking News shelf triggers on

[13] https://www.google.com/search/howsearchworks/mission/users/
[14] For more on Search Quality Raters Guidelines, see www.google.com/search/howsearchworks
[15] See our Search Quality Rater Guidelines: https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

the YouTube homepage automatically when there is a significant news event happening in a specific country and is similarly restricted to authoritative and verified news sources.

- The Breaking and Top News shelves are currently available in more than **40 countries**.
- In 2019, consumption on authoritative news partners' channels on YouTube grew by **more than 60%**.

Left: Breaking News Shelf
Right: Top News Shelf



- **ADDITIONAL PROTECTIONS DURING BREAKING NEWS EVENTS:** On both **Google Search and YouTube**, breaking news events, and the heightened level of interest that they elicit, are magnets for bad behavior by bad actors. Speculation can outrun facts as legitimate news outlets on the ground are still investigating. At the same time, bad actors are publishing content on forums and social media with the intent to mislead and capture people's attention as they rush to find trusted information. To reduce the visibility of this type of content, we

have designed our systems to emphasize authority more while a crisis is developing.

Another way we connect users to authoritative content is by providing contextual information that can be used to help them determine for themselves the trustworthiness of the content they are provided. This isn't possible everywhere, but where we have it, these features let users dig deeper on a story or piece of content.

For example:

- On Google and YouTube, Knowledge and Information Panels may appear in search results to provide context and basic information about people, places, or things that Google knows about.
- On YouTube, for channels operated by broadcasters that are funded or operated by their country's governments, an Information Panel under each video from that channel clearly indicates that the channel receives government or public funding.
- On YouTube, there have been billions of impressions of Information Panels around the world since June 2018.
- On Google News, we provide a 'Full Coverage' button under links to individual articles so that users who seek to explore a story further can easily access a non-personalized, comprehensive set of articles published on the topic. This often features a timeline, tweets, or fact-checks that help further contextualize the story.
- On Google Search and News, we have highlighted fact checks for almost three years as a way to help people make more informed judgments about the content they encounter online. People come across these fact checks billions of times per year, and we have been expanding similar features to YouTube and Google Image Search.

How we work to avoid personal bias in our ranking system and beyond

We build our products for everyone. While our more than 100,000 employees around the world hold a wide variety of views, we have safeguards in place to ensure that we design and enforce our policies in a way that is free from improper bias.

As mentioned earlier, to ensure Search algorithms meet high standards of relevance and quality, we have a rigorous process that involves both live tests and feedback from thousands of trained external Search Quality Raters from around the world.[16]

The Search Quality Rater Guidelines that define the goals of our ranking systems include the criteria that our raters use to assess the expertise, authority, and trustworthiness of pages.[17] These criteria do not include political ideology and specifically provide guidance for raters that *"ratings should be based on the instructions and examples given in these guidelines. Ratings should not be based on your personal opinions, preferences, religious beliefs, or political views."* Furthermore, whether a business, individual, or organization buys ads is not a factor in our search algorithms. We never provide special treatment to advertisers in how our search algorithms rank their websites, nor how our policies are enforced, and nobody can pay us to do so.

In addition, we conduct live traffic experiments to measure how users interact with a new feature before releasing it more widely. Results from these experiments are reviewed by experienced engineers and search analysts. They collectively determine whether the change is approved to launch. In 2019, we conducted over **460,000 experiments** with trained external Search Quality Raters and live tests, which resulted in more than **3,600 improvements** to Google Search.

This commitment goes beyond ranking. A diverse set of external and internal stakeholders are consulted during policy development. Our process involves multiple Google teams, and leaders are involved in finalizing a new or updated policy. We outline our product policies and guidelines in help centers and other fora so that our users can understand the rules that apply to our products.

In addition, we enforce our policies consistently, regardless of who or what is involved. "Gray area" cases – those that approach a policy boundary – are reviewed by multiple people to ensure that an appropriate decision is made, and we have a rigorous quality assurance process for all cases across our products. We approach with similar caution the development and use of the safety lists that help us ensure, for instance, that a website we demonetised for the most

[16] https://www.google.com/search/howsearchworks/mission/users/
[17] For more on Search Quality Raters Guidelines, see www.google.com/search/howsearchworks

severe infringement of our advertising policies is not inadvertently offered the possibility to monetize again via another of our services.

This of course does not mean that our products are 'neutral.' Any ranking inherently involves classification by reference to a specific set of goals or factors. However, we make sure that we publicly document the kinds of goals and factors our products optimize for, and we welcome feedback. For instance, our Search Quality Rater Guidelines outline how we characterize expertise, authoritativeness, or trust for Google Search.

Sometimes, there are no expert, authoritative sources we can elevate in response to a search or perceived user intent. This class of situations, where the data available to respond to a user's query is limited, non-existent, or deeply problematic, have been referred to by researchers as "data voids."[18] We have made progress addressing such data voids during breaking news events. The ranking systems on Google Search and YouTube are trained to detect breaking news events and emphasize authority in search results while a crisis is developing. We continue to explore other ways of addressing the issue of "data voids" across our products and services.

Furthermore, users may decide to seek and select content that our signals determine to be of low-quality, which still appears on our platforms because it does not infringe on our policies. If and when they do, as stated earlier in this paper, we believe it is of fundamental importance to respect their choices.

[18] Data Voids: Where Missing Data Can Easily Be Exploited – Danah Boyd, Michael Golbiewski, 2019 – https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

Reduce: Limiting the reach of borderline content

We set a high bar for amplifying content on our platforms. While every piece of content that is available on our services should be discoverable if users are actively looking for it, not all content is appropriate to recommend to a user – and we have no obligation to do so. That is why our systems and policies seek to ensure that we do not proactively expose users to content that is potentially harmful.

This applies to the features of our products and services where we recommend content to users (e.g., YouTube's recommendations feature), where we give prominent treatment to a piece of content (e.g., Featured Snippets in Google Search results), where content is determined by partnerships or curation (e.g., Knowledge Panels in Google Search or Information Panels on YouTube), and where we may help people complete an intended search based on real searches that happen on Google (e.g., Autocomplete).

Reducing recommendations of borderline content and harmful misinformation on YouTube

In January 2019, we announced that we would begin reducing recommendations on YouTube of borderline content or videos that could misinform users in harmful ways. We continue to extend these efforts to more countries outside the United States, including the United Kingdom, Ireland, South Africa, and other English-language markets. In addition, we have begun expanding this effort to non-English-language markets, starting with Brazil, France, Germany, Mexico, and Spain. We have launched over 30 different changes to our recommendations systems on YouTube in order to reduce recommendations of borderline content and harmful misinformation. In 2019, we saw a more than 70% average drop in "Watch time" of this content coming from non-subscribed recommendations in the United States.

Determining what is harmful misinformation or borderline content is challenging, especially given the wide variety of videos uploaded to YouTube. To do it, we rely on external evaluators from around the world to provide input on the quality of a set of videos. These

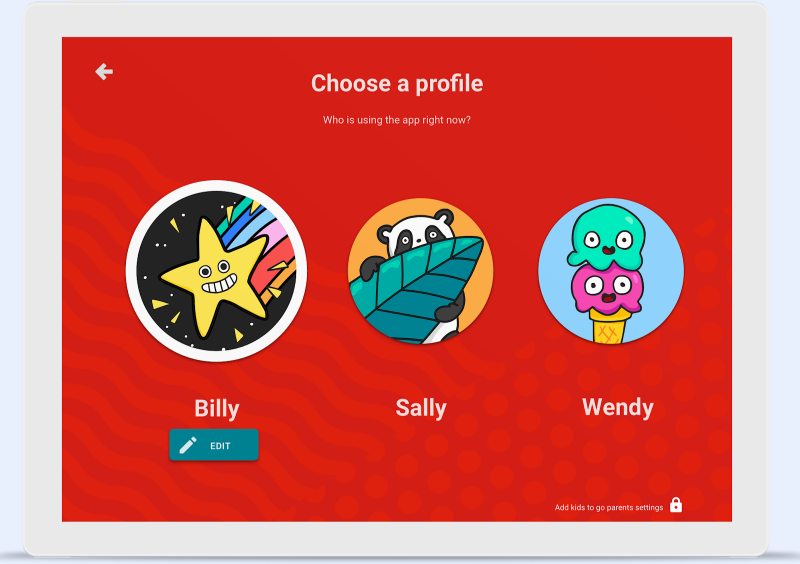
evaluators use the same rater guidelines as Google Search to guide their work. Each evaluated video receives up to nine different ratings, with some content requiring ratings from certified experts in the field. For example, medical doctors provide guidance on the validity of videos about specific medical treatments to limit the spread of medical misinformation. Based on consensus input from these raters, we use well-tested machine learning systems to build models that help review hundreds of thousands of hours of videos every day to identify and limit the spread of borderline content. The accuracy of these systems continues to improve over time.

Building safer experiences for kids and families

We build products for kids and families from the ground up to help parents and educators support safer experiences for their children and students.

- **Family Link** is available by default on the latest Android operating system and helps parents stay in-the-loop as their child explores the internet on a compatible device. The app lets parents set digital ground rules for their family, like managing the apps their child can use, keeping an eye on screen time, or setting a bedtime and daily limit for their child's device.
- **Assistant for Families**: Children signed in with their own account, created through Family Link, are given a more kid-friendly experience using the Google Assistant – with access to answers and features created especially for them.
- **YouTube Kids** provides a separate YouTube experience designed especially for children, which parents can control. The app uses a mix of filters, user feedback, and content moderators to keep the videos in YouTube Kids family-friendly, allowing children to explore a catalog of content in a safer environment. In addition, parental control tools allow families to hand-select all of the content their children watch, or to choose content from third-party collections assembled by experts, like UNICEF and PBS Kids.

YouTube Kids



- **Google Play** has designed policies aimed to ensure that apps for children have appropriate content, show suitable ads, and handle personally identifiable information (PII) correctly. They also reduce the chance that apps not intended for children could unintentionally attract them. We are now asking every developer to thoughtfully consider whether children are part of their target audience via our new policy requirements, and requesting that in-app ads served to children are from an ads network that has certified compliance.
- **Expedition** and **Socratic** support kids' classroom and learning experiences.

For more information on our work, visit https://safety.google/families/

Reward: Setting a higher standard for monetization

We set a particularly high bar for information quality on services that involve advertising and content monetization, which includes Google Ads and AdSense. We have no desire to derive revenue for ourselves, or any other business, from harmful content or behaviors. In addition, given that many bad actors seek to make money by spreading harmful content, raising the bar for monetization can also diminish their incentives to misuse our services.

We prohibit hateful content and deceptive behavior on our advertising products. This includes prohibiting publishers that seek to use our services from displaying ads on pages aimed at harassing and bullying, or otherwise promoting dangerous or derogatory content. We also prohibit publishers that seek to misrepresent the primary purpose of their web destination. For instance, in 2017, we discovered that a group of publishers in Macedonia had created a group of websites all presenting themselves as American news outlets. Upon investigation of that group, it became clear that the sites were not legitimate news outlets and we demonetized them.

We also restrict certain kinds of businesses from using our advertising products in order to prevent users from being exploited. For example, studies show that for-profit bail bond providers in the United States make most of their revenue from communities of color and low-income neighborhoods when they are at their most vulnerable, including through opaque financing offers that can keep people in debt for months or years. After working with experts in this space, we decided to take action to protect our users by creating a new policy to restrict ads that promote bail bond services on our platforms.

On YouTube, our partner program allows creators to monetize their content and access additional tools to build their channels. Creators must meet a threshold of subscribers and public Watch time and follow YouTube's monetization policies. We closely monitor signals like community strikes, spam, and other abuse flags. Violations of these policies may result in disabling ads from certain videos, disabling a channel's AdSense account, suspension from the YouTube Partner Program, or channel termination.

Preventing the monetization of low-quality content

In addition to it not being consistent with our company purpose, it is also not in our business interest to allow for the advertising or monetization of low-quality content. Advertisers typically prefer not to profit from this sort of content, or enable it for monetization, and we have a vested interest in ensuring that they view us as trustworthy partners in protecting the integrity of their brands. We invest significantly in human and technology resources to prevent it. Neither we, nor the advertisers that rely on our platforms, wish to be associated with such low-quality content.

For instance, In 2018:

- We terminated nearly **734,000 publishers and app developers** from our ad network,
- We removed ads completely from nearly **1.5 million apps**.
- We also modified the YouTube Partner Program eligibility requirement for monetization to **4,000 hours of Watch time** within the past 12 months and **1,000 subscribers**.

And in 2019, we **terminated over 1.2 million accounts and removed ads from over 21 million web pages** that are part of our publisher network for violating our policies.

We continue to improve our policies to ensure that these and other services are not used to create or propagate harmful information experiences that lead to monetization for anyone, especially Google.

Working with others

Managing information quality and content moderation across our products and services requires significant resources and effort. The speed at which content is created and shared, and the sophisticated efforts of bad actors who wish to cause harm, compound the challenge of fulfilling our mission. Fortunately, we are not alone.

We work with many talented experts and organizations across the technology industry, government, and civil society to ensure that we are doing everything we can to set good policies, establish, share, and learn from industry best practices, as well as get ahead of emerging challenges. Here are some examples of that work.

Collaboratively identifying violative content

Building on our own efforts, we rely on a community of partners who have specific subject-matter expertise to help us identify content that violates our rules of the road.

The YouTube Trusted Flagger program was developed by YouTube to help provide robust tools for individuals, government agencies, and non-governmental organizations (NGOs) that are particularly effective at notifying YouTube of content that violates our Community Guidelines.[19] The program provides these partners with a bulk-flagging tool and provides a channel for ongoing discussion and feedback about YouTube's approach to various content areas.

The program is part of a network of **over 180 academics, government partners, and NGOs** that bring valuable expertise to our enforcement systems. For instance, to help address violent extremism, these partners include 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.

Participants in the Trusted Flagger program receive training in enforcing YouTube's Community Guidelines. Because their flags have a higher action rate than the average user, we prioritize them for review. Content flagged by Trusted Flaggers is subject to the same policies as content flagged by any other user and is reviewed by our

[19] https://support.google.com/youtube/answer/7554338?&ref_topic=2803138

teams who are trained to make decisions on whether content violates our Community Guidelines and should be removed.

We also commission or partner with organizations specialized in tracking and documenting the work of threat actors who seek to target our products and services around the world. We typically do not share much information about these partnerships in order to protect these companies and their employees from the threat actors they monitor. Some examples of this work are public, such as our work with FireEye, a cybersecurity company, to detect a number of security incidents and influence-operations.

Evolving and improving our policies

We must always work to improve our policies in light of changing user behaviors or expectations, and in response to the constantly evolving tactics of malicious actors. We often seek the advice of subject-matter experts in the appropriate field, gaining important perspectives from academic researchers, civil society organizations, and others in the industry.

Hate speech policies on YouTube

After consulting with dozens of experts in subjects like violent extremism, supremacism, civil rights, and free speech, we updated our policies on YouTube to prohibit videos which allege that a group is superior in order to justify discrimination, segregation, or exclusion based on qualities like age, gender, race, caste, religion, sexual orientation, or veteran status.

We also worked with experts to develop more stringent harassment policies, and, as a result of these changes and our ongoing enforcement, we removed over 100,000 videos and 100 million comments for hate and harassment in the first quarter of 2020 alone. That said, we know there's more work to do, and we continue to examine how our policies and products are working for everyone.

Developing best practices to improve the internet ecosystem

We work with other technology companies and industry partners to address challenges that span multiple products and ecosystems by identifying where cooperation would be beneficial and where the resources of a company like

Google can help increase the capacity of others. This type of collaboration is often the most effective mechanism for fighting bad actors at scale.

Fighting child sexual abuse material

In order to help eradicate the horrors of Child Sexual Abuse Material (CSAM), Google joined with other industry members in the Technology Coalition in 2006. We make cutting-edge technology available to qualifying industry and non-governmental organizations for free in order to help identify, remove, and report illegal CSAM more quickly and at a greater scale. In the last decade-plus, member companies have made progress with the development and roll-out of innovative technology to combat CSAM, and, in 2020, the Coalition announced 'Project Protect,' a renewed investment and strategic plan to enhance our collective work.[20]

Tools like CSAI Match and Content Safety API, which were developed by Google and YouTube engineers, help prioritize potentially illegal content for review while identifying known and never-before-seen CSAM. In addition to being used on our platforms, these tools are also being used by companies like Adobe, Tumblr, and Reddit to aid in the faster identification of potential victims of CSAM, while reducing the toll on content moderators.

Countering terrorism content

In order to substantially disrupt terrorists' ability to promote terrorism, disseminate violent extremist propaganda, and exploit or glorify real-world acts of violence using our platforms, we have partnered with others in the industry to establish the Global Internet Forum to Counter Terrorism (GIFCT).

GIFCT
Global Internet Forum to Counter Terrorism

Among other important initiatives, GIFCT allows participating companies and organizations to submit hashes, or 'digital fingerprints,' of identified terrorist

[20] https://www.technologycoalition.org/2020/05/28/a-plan-to-combat-online-child-sexual-abuse/

and violent extremist content to a database so that it can be swiftly removed from all participating platforms.

By sharing best practices and collaborating on cross-platform tools we have been able to:

- Increase our hash-sharing database to 200,000 hashes.
- Build a global research network that aims to better understand the ways in which terrorists use technology.
- Bring new members to the GIFCT and engage more than 100 smaller technology companies through workshops around the world.

Partnering to fight terrorism

This is a quickly evolving challenge that requires us to continue improving our tools and approach alongside the threats we face. GIFCT is a crucial part of this. For example, in 2019, the tragic events of Christchurch underscored the urgent need to improve the exchange of information between platforms to address the challenge posed by live uploads and coordinated reuploads of terrorist content.

We were proud to be part of the Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content Online and to make progress toward its commitments. For instance, GIFCT developed and implemented a protocol for responding to real-world events involving the murder of defenseless innocents and civilians, which has since been activated to address a violent attack featuring perpetrator-filmed content. At the end of 2019, YouTube co-organized a crisis prevention workshop with the New Zealand Government in Wellington, NZ, attended by participants from around the globe, during which the protocols were further refined.

In 2020, GIFCT is growing into an independent organization, led by an executive director and supported by dedicated technology, counterterrorism, and operations teams. This new, independent GIFCT continues to support a program of knowledge-sharing, technical innovation, and shared research in collaboration with experts, civil society, and government, building on lessons from 2019. It also continues its progress in fulfilling the commitments of the Christchurch Call to Action, supporting academic research, promoting counterspeech efforts online, and empowering a broad range of technology companies to prevent and respond to abuse of their platforms.

We continue to develop and learn from these collaborations over time and seek more opportunities to develop best practices jointly with partners in industry and government.

Supporting information quality through regulation

Thoughtful regulation is good for society and the internet, and nowhere is it more important to get the balance right than in the debate over content online.

Many laws, from consumer protection to defamation to privacy, already govern content online. A smart legal framework for online platforms has been essential to enabling a reasonable approach to illegal content. For instance, appropriate safe harbors spell out how online platforms can fulfill their legal responsibilities when notified about illegal content, and ensure an online platform that takes other voluntary steps to address illegal or otherwise harmful content is not penalized. These laws have promoted the free flow of information, innovation, and economic growth, while giving platforms the legal certainty they need to combat problematic content.

Effective oversight of content moderation practices can also play a complementary role. Throughout the internet's history, industries, policymakers, and civil society have worked on codes of practice to guide appropriate behavior by online services. As content sharing services like social media and video sharing sites have become more important to public discourse, oversight methods will continue to evolve as well, so as to better review platforms' efforts in light of best practices.

We think new forms of oversight can work well when they focus on a specific, clearly defined problem and do three things:

- **Set out standards for transparency and best practices:** Transparency provides the starting point for effective practices and the basis for an informed discussion. Because technology is not static and new forms of communication continue to evolve, oversight should take a flexible, collaborative approach that supports best practices, and promotes research and innovation.
- **Address systemic, recurring failures, not one-offs:** The scope and complexity of modern platforms requires an approach that focuses on overall results rather than anecdotes. While we will never eliminate all problematic content, we should recognize progress in making that content less prominent and use data-driven approaches to understand whether particular errors are outliers or representative of a more significant problem.
- **Foster international cooperation:** Given the multinational nature of modern platforms, and recognizing people's abilities to communicate and access information from other people across the world, countries

should share best practices with one another and avoid conflicting approaches that impose undue compliance burdens. International coordination should strive to align on broad principles and practices. That said, individual countries will make different choices about permissible speech based on their legal traditions, history, and values, consistent with international human rights obligations. Content that is illegal in one country may be lawful in another, and no one country should be able to impose its rules on the citizens of another country.

Responding to misinformation about the COVID - 19 Pandemic

We rely on the principles and levers outlined above to address all new developments and challenges that relate to information quality or content moderation across our services. The coronavirus pandemic has been one such development – unexpected and unprecedented in its magnitude.

To address it, we carefully examined our policies and practices to ensure we were addressing emerging issues. For instance, as we found that COVID-19 was becoming a lure for scams of various sorts, we dedicated microsite to help users identify and protect themselves from COVID-19-related scams.[21]

One particularly pertinent area of our work for purposes of this paper is how built upon our prior work against dis- and misinformation[22] in order to make sure that we elevate authoritative health information and that we combat harmful medical misinformation across our services.

Elevating trustworthy information around COVID-19

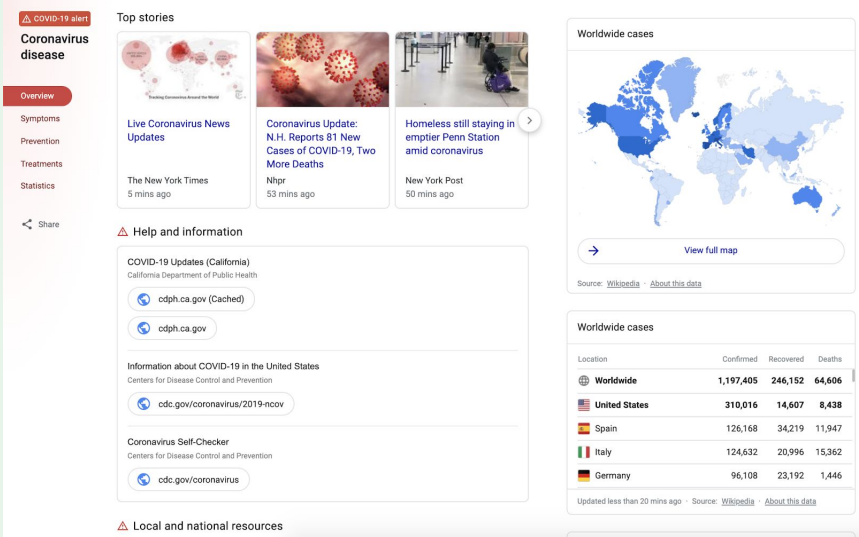
We have worked to surface trusted information and partner with health organizations and governments in order to bring our users authoritative information in a rapidly changing environment:

- **In Search**, we have introduced a comprehensive experience for COVID-19 that provides easy access to information from health authorities alongside new data and visualizations. This new format organized the search results page to help people easily navigate resources and makes it possible to add more information as it becomes available over time. This experience came as a complement to pre-existing work on Google Search and Google

[21] https://safety.google/securitytips-covid19/
[22] See for instance our February 2019 white paper, "How Google Fights Disinformation": https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Disinformation.pdf

Dedicated Search Experience, available internationally

News to recognize sensitive events and contexts, and our systems are designed to elevate authoritative sources for those classes of queries.



- **On the Google HomePage,** in partnership with the World Health Organization and other health authorities, we have promoted important guidance to prevent the spread of COVID-19. The efforts, including "Stay Home" doodles and messaging on our homepage, have launched in more than 100 countries to date.

- **Across YouTube**, we elevated authoritative sources such as the WHO and local authorities to help users get the latest COVID-19 information. We've launched a COVID-19 news shelf on our homepage that features stories from authoritative publishers and local health authorities, health information panels in search results that feature information on COVID-19 symptoms, prevention, and treatment, and links to local health authorities on the watch pages of COVID-19 related videos. In addition, YouTube elevates content from authoritative channels such as news organizations or health authorities when our systems detect that a user's search is health-related.

- **In Google News**, we have created a new COVID-19 section with links to up-to-date, relevant stories from the international to local levels from a variety of authoritative sources. The section is now available to users across 40 top impacted markets and puts local news front and center by highlighting stories about the virus from local publishers in the reader's area.

- On **Google Maps**, we have made it easier to find authoritative information about local health resources, including COVID-19 testing sites, shelters, food banks and virtual healthcare options where available. We also used authoritative data sources to display updated information about whether local businesses are open during COVID-19. In addition, we have provided businesses with new ways to update their listing information and service offerings such as restaurants that are offering takeout or delivery, but are closed for dine-in.

- On **Google Play**, we prioritized the review and publication of policy-compliant apps published, commissioned or authorized by official government entities and public health organizations. Authorized COVID-19 apps must comply with all Play Developer policies, including User Data, Permissions, and Malicious Behavior. We also launched a "stay informed" page in the Play Store with apps that can help users stay informed and prepared during the crisis, using authoritative sources such as the WHO app.

- A **new website**, which provided resources dedicated to COVID-19 education and prevention, has also been released. As of the release of this paper, it continues to be available on www.google.com/COVID-19 in more than twenty languages and we're continually working to expand its coverage.

Combating health misinformation across our services:
In addition to elevating authoritative information, we have taken active steps to detect and remove COVID-19 related misinformation that contradicts guidance from health authorities and may result in real-world harm:

- On **YouTube,** our Community Guidelines prohibit content that encourages dangerous or illegal activities that risk serious physical harm or death, including certain types of medical misinformation. As the COVID-19 situation has evolved, we have partnered closely with the World Health Organization and local health authorities to ensure that our policy enforcement is effective in preventing the spread of harmful misinformation relating to COVID-19. Our policies prohibit, for example, content that explicitly disputes the efficacy of WHO or local health authority advice regarding social distancing that may lead people to act against that guidance. We enforce these policies diligently and, in addition, continue the work we initiated in 2019 to reduce recommendations of borderline content or videos that could misinform users in harmful ways.

- On **Google Play**, our policies prohibit developers from capitalizing on sensitive events. Our long-standing content policies strictly prohibit apps that feature health-related content or functionalities that are misleading or potentially harmful, including about COVID-19. Apps that violate these policies will be removed.

- On **Maps**, our policies prohibit misinformation about prevention, transmission and treatment services, as well as allegations that an individual contracted COVID-19 at a particular location. These types of contributed content will be removed.

- On **Google Ads**, our policies do not allow ads that potentially capitalize on or lack reasonable sensitivity towards a sensitive event, such as a public health emergency. Over time, we started phasing in allowances for COVID-related ads from government organizations, healthcare providers, non-governmental organizations, intergovernmental organizations, verified election ads advertisers and managed private sector accounts with a history of policy compliance who want to get relevant information out to the public. Ads that were allowed still had to abide by our policies, which also disallow the promotion of harmful medical or health claims and practices. In addition, we enforced a temporary restriction on personal protective equipment and we are taking additional steps to prevent artificially inflated prices that limit or prohibit access to other essential items on Google's network. More information can be found in our Google Ads Help Center.

Over the course of the pandemic, we have continuously reviewed and improved these policies and our enforcement in order to respond to the changing landscape of COVID-19 related misinformation.

Supporting content moderators:
In the face of temporary reductions in our extended workforce, we reallocated employees to prioritize addressing egregious content and supported their doing this work onsite, taking extra precautions on hygiene and providing private transportation. These content moderators ensured we still had capacity to action high priority workflows and flags for egregious content, including flags from our Trusted Flagger program and governments.

Where feasible, we relied more heavily on automated systems, to reduce the need for people to come into the office. Given the resulting risk of false positives (more legitimate content being automatically but incorrectly removed), we also worked to ensure content creators

could appeal and would not wrongly receive strikes against their accounts.

Helping users stay abreast of our work:
Over the course of the crisis, we have provided publicly available information so as to help users, civil society, and other interested parties abreast of our work via dedicated pages on Google's blog and on YouTube's Help Center.

Conclusion

Information quality and content moderation are crucial to Google's mission. They embody a commitment to our users to provide trustworthy, useful information that meets their needs and protects them from harm. It is a commitment on which we are judged every day and in every user interaction.

They are also unique, significant challenges. Reasonable people can disagree on desirable outcomes in addressing them. Bad actors are persistent and creative. And the breadth of content available online makes it impossible to give each piece of content an equal amount of attention, human review, and deliberation. Our work to address them will not soon be complete.

Never in our short history has the impact of our work mattered more to society. We continue to invest in developing and improving the policies, products, tools, processes, and teams that handle information quality and content moderation across our platforms. It is critical to our business and to the societies in which we operate that we get it right.

We are confident and optimistic that the approach we have described in this paper is meeting the challenge with measurable success. This approach to supporting information quality across our various products and services will continue to adapt to the changing nature of the challenge, using these four complementary levers:

- **Remove**: we set responsible rules for each of our products and services and take action against content and behaviors that infringe on them. We also comply with legal obligations requiring the removal of content.
- **Raise**: we elevate high-quality content and authoritative sources where it matters most.
- **Reduce**: we reduce the spread of potentially harmful information where we feature or recommend content.
- **Reward**: we set a high standard of quality and reliability for publishers and content creators who would like to monetize or advertise their content.

We will continue exercising these four levers across our products and services, learning and improving over time, while collaborating with industry partners, civil society, and governments. Together, we will chart a future that preserves users' rights and upholds the values of the open web while providing a better and safer experience for everyone. These efforts are in direct alignment with Google's founding mission, with the demands of our users, and the societies in which we operate.

We welcome your feedback on this approach and our progress.