

Your response

Please refer to the sub-questions or prompts in the annex of our [call for evidence](#).

Question	Your response
<p>Question 1: Please provide a description introducing your organisation, service or interest in Online Safety.</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>About Google</p> <p>Google’s mission is to organise the world’s information and to make it universally accessible and useful. Google achieves this by providing users with a range of services to exchange information and ideas. We believe deeply in technology’s ability to facilitate expression and access to knowledge, but we also understand the responsibility we have to keep our users safe. We put safety at the heart of how we develop our services — and our investment and innovation have often put us at the forefront of positive industry change in this area. We recognise that new risks are continually emerging and we are always thinking about what more can be done to protect users. We look forward to working with Ofcom as we further develop our approach.</p> <p>Google’s interest in online safety</p> <p>Safety is core to how Google develops and operates its services, and we understand our responsibility to keep users safe while protecting their privacy and promoting the free flow of information. Promoting access to trustworthy information and effective content moderation are crucial to Google’s mission. They embody a commitment to our users to provide useful information that meets their needs and protects them from harm. 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.</p> <p>We believe that thoughtful regulation is good for society, business and the internet. The Bill’s goal to strengthen online safety whilst protecting freedom of expression requires a careful balance to be struck and it is important to consider in this regard the outcomes this Bill aims to achieve. Our view is that thoughtful regulation should drive up safety standards for users, but it must also mitigate disproportionate and unintended impacts on freedom of expression, privacy, and innovation which benefits users and drives economic growth.</p> <p>In our view, this is an opportunity for Ofcom to establish a genuinely world-leading approach that supports the UK’s ambitions to be both the safest place to be online and a hub for tech investment and innovation. We recognise that Ofcom will be entrusted with difficult decisions that balance safety, freedom of expression, privacy, and innovation. We welcome Ofcom taking a nuanced approach to these decisions that draws on the</p>

available evidence. We will support this process by providing evidence that draws on our own extensive experience of navigating content issues online.

Google's services¹

Each of the products and services that Google offers has a different purpose, and their users have different expectations of the kind of content they will encounter on each, and whether they will encounter other users. Google **Search** serves as an index of all pages available on the open web - users expect to find search results reflecting every relevant webpage concerning their query and they do not interact with other users. **YouTube** is a platform for uploading and sharing content as part of a community, and is based on user-to-user interactions. Other services are designed for a specific purpose (such as navigation on Maps), but may include some ancillary user-to-user functionalities (such as reviews of locations on Maps). By way of additional detail:

- **YouTube:** YouTube is an online video sharing platform that allows users to create, share, view and comment on user-generated content. Users can search for and watch videos, create a personal YouTube channel, like/comment/share others' videos, subscribe to follow other YouTube channels, and create playlists to organise videos.
- **Search:** Google Search processes billions of searches per day. Google uses automation to discover content from across the web and other sources. Search functions by crawling the web, following hyperlinks from page to page, and creating an index of the web pages it finds. This is fundamentally different to "user-to-user services": search engines are essentially indexes of the entire web, allowing users to access trillions of web pages through search results; user-to-user services allow users to upload and share content. When an individual enters a search query, it uses algorithms to return search results linking to the relevant web pages in the index, ranked from most to least relevant based on over 200 factors. Search results and their rankings may differ from one search to another based on several factors, including location, time, and in some cases the user's search history.
- **Maps:** Google Maps is an online-based consumer map and navigation service. The core feature and primary content is the underlying maps, images and data facilitating navigation. To enhance the map and navigation service, Maps also

¹ Most of our services (including Search and YouTube) are provided to users in the UK by Google LLC. This response relates to those services as provided by Google LLC and is focused primarily on YouTube and Google Search because of the reach of these services as well as the types and range of content that can be accessed through them. We should note that we understand that the fraudulent advertisement duties are not within the scope of this Call for Evidence. We have therefore not addressed Google Ads in this response. We would be happy to discuss separately with Ofcom our approach to addressing fraudulent advertising on Google Ads.

allows users to share reviews, photos, and videos on places (including businesses).

- **Workspace:** This broad product category includes Google Drive (where files can be stored and shared via link), user-to-user video products (like Google Meet) and user-to-user messaging products (like Google Chat).
- **Photos:** Photos is a service that allows users to manage, edit, store, and share their photos and videos. Photos and videos can be shared through the service or via link.
- **Assistant:** Google Assistant allows users to use voice commands to search the web on mobile and home automation devices.

Google revenues and business model

In terms of our global and UK revenues:

- Alphabet Group revenue was USD 257.6 bn for calendar year ended 2021².
- Google UK Limited³ revenue was GBP 1.8 bn for year ended 30 June 2020⁴.

Across the company, Google's main source of revenue is advertising – mostly from ads on the sites and apps of our own products and services. By serving ads, we can keep many Google services open to anyone with an internet connection, no matter where they live or what their background is, free of charge (see more detail [here](#)).

Google's approach to online safety

Google takes a comprehensive approach to online safety, involving large global teams and experts with a deep understanding of the key issues. In our experience, this is the best way to avoid over-relying on any single tool and ensures we can address the different aspects that contribute to a safer online experience. Our approach has three elements: strong policies and guidelines, technological innovation and enforcement, and working in partnership with others:

² As Alphabet is the parent entity of Google's broader corporate group, Alphabet's total revenues are submitted. The revenue data is sourced from Google's accounting systems. These represent booked revenues which are invoiced to customers (advertisers, developers and publishers), and reflect certain manual accounting adjustments (such as sales incentives and invoicing adjustments due to, for example, invalid clicks for Search (i.e., spam)). Revenue data is converted from local currency to USD based on a monthly average rate.

³ Google UK Limited is a wholly-owned subsidiary of Google LLC. Google UK Limited does not provide any online services and is not the contracting entity to the terms of service governing the use of Google's consumer services.

⁴ The data reflects revenue recognised in the UK based on the IFRS. Google UK Limited has undergone an accounting period realignment. The next set of its financial statements will be filed by 30 September 2022 covering the 18-month period 1 July 2020 to 31 December 2021.

- **Policies and guidelines:** Google has clear and publicly available product-specific “policies” and “community guidelines” developed in partnership with experts (please see, for example, our response to [Question 6](#)). These are the “rules of the road” and are intended to make sure that all users have a clear understanding of acceptable and unacceptable content, and online behaviour. They also explain the process by which a piece of content, or its creator, may be removed from the service.
- **Innovation and enforcement:** Safety is incorporated into the design of each of our services and we constantly iterate and improve. We have built a range of products, tools and approaches across our different services that ensure users can have a safe experience. For example, YouTube Kids provides a separate YouTube experience designed especially for children that parents can customise. Our policies and guidelines are robustly enforced by these systems and tools, and we describe them in greater detail in our response to [Question 6](#) below.
- **Partnership:** Whilst fully accepting our own responsibility to keep users safe on our own services, we also believe that making the UK the safest place to be online requires a whole-of-society approach. This is why we have established long standing partnerships with experts in areas such as hate speech and media literacy. Media literacy and digital citizenship is vital for helping people make the most of online opportunities while empowering them to protect themselves and their families from the potential risks. Our work here includes the “[Be Internet Legends](#)” online safety learning programme that we deliver in partnership with ParentZone (as explained in our response to [Question 20](#)), which helps children to be confident and safe explorers of the online world. We have provided online safety training to over five million primary school children in the UK.

Google supports the need to ensure that online services implement effective systems and processes to provide user safety. Modern platforms are complex, requiring a systematic approach focused on prevention rather than on one-off content moderation decisions.

Ofcom’s approach

Google welcomes the opportunity to engage with Ofcom on how to strengthen online safety in the UK, and we welcome the methodical and consultative approach Ofcom is taking. We would find further clarity particularly welcome in the following areas:

- **Ongoing engagement with providers:** We welcome the provisions in the Bill for Ofcom to consult with industry. We have worked with policy makers and regulators for a number of years to inform regulatory processes. We would be

happy to arrange technical workshops with our expert teams so that the Ofcom online safety teams drafting the Codes of Practice can benefit from industry technical expertise and experience. We would be happy to work with Ofcom to find workable solutions. In particular, we would appreciate the opportunity to work with Ofcom during the development of guidance on “Notices to Deal” as well as before each decision to “accredit” technology.

- **Requirements to use technology and the implications for freedom of expression:** We would welcome clarity in the Codes of Practice on the fundamental challenges of addressing the Bill’s safety duties alongside the duties to protect users’ rights to freedom of expression. We have concerns that, without further guidance, potential requirements to use technology to monitor content and the ambiguous language of the safety duties could have profound implications for our approach to content moderation and lead to the over removal of lawful content. It is unclear whether, in its current form, the Bill would require us to calibrate tools to remove content even when uncertain of its illegality, which would significantly risk freedom of expression. We would welcome opportunities to share these views and to receive further guidance from Ofcom on how the safety duties can be applied in a manner that does not risk over-blocking of legitimate content.
- **Mitigating risks to users’ privacy:** We welcome Ofcom’s ongoing commitment to work closely with the Information Commissioner’s Office (ICO), including through the Digital Regulation Cooperation Forum (DRCF), on the interaction between privacy and online safety law. We welcome continued guidance from, and cooperation between, both regulators on how to appropriately navigate the protection of user privacy and online safety simultaneously, for example, on how we effectively safeguard user privacy when collecting and processing data to identify a potential grooming case. In particular, we would appreciate clarity from Ofcom and the ICO on how to comply with privacy obligations while using automated technology to proactively monitor content. Additionally, requirements to use identity verification and age assurance measures could, without further guidance and clarity, lead to excessive collection of users’ data, including children’s data.
- **Proportionate and differentiated approach for Search services:** As the Joint Parliamentary Committee on the draft Bill concluded, because search engines operate differently from social media, **“the codes of practice drawn up by Ofcom will need to recognise the specific circumstances of search engines to meet Ofcom’s duties on proportionality”**. Search engines’ unique and valuable role in providing access to information (rather than hosting content) means there should be more emphasis on providing tools that empower users to have a safe search experience, rather than on the removal of search results. We

would value further clarity through the Codes of Practice on measures that are technically feasible and appropriate to the particular functionality of search engines; as mentioned above, we would be happy to arrange technical workshops with our expert teams.

- **Clarity on treatment of journalistic content:** The Bill requires a special expedited complaints procedure to be made available in respect of decisions to restrict or remove “journalistic content”, which is defined not just as news publisher content but also user-generated content which is “*generated for the purposes of journalism*”. We would welcome the opportunity to work with Ofcom to ensure that the detailed requirements relating to this duty remain workable and not open to abuse by bad actors seeking to spread harmful content under the guise of citizen journalism. Additionally, any “must not remove obligation” for news publisher content must be implemented carefully to ensure platforms, including YouTube, are not required to expose users to content they have deemed to be illegal or harmful. For example, Ofcom research recognises that warning labels for graphic content are effective for mitigating harm; we would be concerned that a “must not remove” obligation could prevent the use of such features. We discuss this in greater detail in our response to [Question 11](#).

Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services?

Is this response confidential? – Y / N (delete as appropriate)

By way of context and background, it is important to note that we have two distinct, internal processes for the removal of content:

(i) **Policy violations:** we remove content that violates the policy of a particular service. This process addresses a broad range of content, which is wider than the Bill’s definition of “illegal content”, and covers content ranging from illegal online child sexual abuse material (CSAM) and fraud, to legal but potentially harmful content like spam and misinformation. We remove content for policy violations based on user reports as well as through our own content moderation processes.

(ii) **Legal removals:** We remove content that our legal analysis has determined to be unlawful under applicable law, in response to a notification from a third party, such as a user or an authority. Examples include defamatory material (on YouTube), material in relation to which we have received a valid “right to be forgotten request” (for Search), or material in relation to which we have received a valid court order. We measure the number of court and government legal removal requests biannually (across all products), and publish this information in [transparency reports](#).

Where we refer to content that is removed under either one of these processes, it is important to note that the content in question may either be “illegal content” for the purposes of the Bill, or may fall outside that definition, because we remove content for a broader variety of reasons than illegality (as defined by the Bill) alone.

Presence or quantity of illegal content on Search

Google **Search** works at the scale of the web. Given the volume of content on the web, the speed at which new content is created, and the need for human review to identify certain types of illegal content (in particular, where legal nuance, competing rights, and context are relevant), it is not currently possible for Google to quantify the illegal content on the open web at any given time.

However, it may be useful to provide some information on the actions that we take in relation to violative and illegal content. Like all search engines, we do not host content on the web and so we cannot remove content from the web; this can only be done by webmasters themselves. However, what we can do is to either “delist” or “demote” content:

- The term “delisting” refers to a process by which we “remove” links to certain web pages from the lists of displayed search results. This stops returning to users links to certain web pages (at times for all search queries, and at times following only certain search queries) and thereby prevents those web pages from being accessed through Search. Content that has been delisted is, however, still accessible via the open web, via direct navigation to hosting sites, social media platforms, and/or via searches on other search engines.
- The term “demoting” (referred to in the Bill as giving content “a lower priority in search results”) refers to the process of ranking links to certain web pages lower in response to certain search queries and thereby makes it less likely that those specific web pages are accessed through the Search service. Conversely, we can prioritise helpful webpages.

Our content policies for Google Search specify that we:

- Delist search results that lead to child sexual abuse imagery or material that appears to victimise, endanger, or otherwise exploit children (we also report CSAM, as explained in our response to [Question 10](#) below).
- Delist certain personal information that creates a significant risk of identity theft, financial fraud, or other specific harm.
- Delist non-consensual explicit imagery (NCEI).

- Delist or demote spam, which we define as results that exhibit deceptive or manipulative behaviour designed to deceive users or game our search systems.

By way of example, between July and December 2021, 580,380 URLs were de-listed from the Search index for containing CSAM.

In addition, we prioritise useful pages when user queries indicate an urgent need for certain kinds of critical safety information. For example, for search queries that might indicate suicidal intent, a results box is displayed at the top of the search results page with the phone number of the Samaritans, who can provide help and support. We discuss this in more detail in our response to [Question 19](#) below.

Presence or quantity of illegal content on YouTube

On **YouTube**, our [Community Guidelines](#) provide a framework for what is and isn't allowed on the platform, and our Terms of Service require that any content that users upload complies with our Community Guidelines and with any applicable laws.

We measure our global enforcement of our Community Guidelines for YouTube and publish this information in quarterly [Community Guidelines Enforcement Reports](#). These reports include charts explaining, by reference to the reason for removal, the number of channels removed and the number of videos removed. These removals comprise content removed both for policy violation and unlawfulness under applicable law (as explained above).

By way of example, between April and June 2022, almost 4.5 million videos were removed from YouTube. The breakdown of reasons for these removals is as follows:

- 1,383,028 million removals in "child safety" category
- 900,014 removals due to "violent or graphic" content
- 666,315 removals due to "nudity or sexual" content
- 533,896 removals due to "harmful or dangerous" content
- 499,719 removals in "harassment and cyberbullying" category
- 150,833 removals in "spam, misleading and scams" category
- 145,688 removals in "hateful or abusive" category
- 122,660 removals in "misinformation" category
- 72,990 removals in "promotion of violence and violent extremism" category

As another example, for CSAM specifically (which is a sub-category of "child safety"), in 2021, YouTube made nearly 250,000 reports to the National Center for Missing and Exploited Children (NCMEC), relating to almost 270,000 individual pieces of content. This process is explained further in our response to [Question 10](#) below. Figures for the first half of 2022 will be published shortly in our [CSAM Transparency Report](#).

In addition to the data that we publish which shows the volume of content (including illegal content) that we have removed from our services, we also publish “[Violative View Rates](#)” (VVRs) in respect of YouTube content. These VVRs show how many times content has been viewed before it is removed for breaching our policies. We see these VVRs as our “North Star” for measuring our progress in combating harmful content and we believe that sharing these with the public is an important way to create accountability. Our analysis shows that, in January to March 2022, of the 3.8 million videos removed from YouTube for violations of Community Guidelines, 33.7% were never viewed and a further 33.4% were viewed under ten times. In Q1 2022, the VVR was 0.09-0.11%, meaning that out of every 10,000 views on YouTube, only 9-11 came from violative content.

How the quantity of illegal content might vary across services with particular users

It is obviously important that effective forms of protection are put in place when it comes to vulnerable users, which is why Google has developed specific tools and products to keep vulnerable users, including children, safe online. These are described in more detail in our response to [Question 24](#) below.

Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content presented by your service?

Is this response confidential? – Y / N (delete as appropriate)

We conduct assessments across our products and services on the risk of harm to our users. How we assess the risk of harm varies by product, in part because harm manifests itself differently depending on the service and context. Whilst a universally recognised harm may be prohibited across all of our products and services, it can appear on each product and service differently. We assess the risk of harm to an individual, harm to a group based on a specific attribute (for example, race, gender, etc.), and harm that may affect an entire society, such as an attempt to interfere with elections or civic processes.

The process by which we create our policies is based on risk assessment. Our four-step process (shown in the diagram below) works as follows.

The first step we take is to **identify emerging harms and gaps in our existing policies**. To do this, we consider expert input, user feedback and regulatory guidance. For example, with YouTube, our Intelligence Desk monitors the news, social media and user reports from around the world to detect new trends. A key part of our approach is to anticipate problems before they emerge. We also 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 engage in conversations with regulators around the world - their perspectives and concerns directly inform our policy creation process.

Policy creation process

Identifying emerging harms and gaps in existing policies

Determine whether the harm has been addressed



Secondly, we **gather as many examples** of how a particular harm has manifested itself on our services, or might manifest itself 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 behaviour harmful, as well as the risks that an overbroad policy would pose.

Thirdly, we **develop draft standards and enforcement guidelines**. We test draft guidelines against the counter-examples to mitigate against the possibility of public interest content being caught up in any policy change. We also consult with multi-disciplinary experts both inside and outside of Google. We then work to resolve any conflicts thrown up by this process and ensure the new guidelines are coherent. Finally, we test policies until we are confident that we can ensure they can be consistently applied, before rolling them out further.

Fourthly, before we begin implementation and enforcement of a new policy, we **determine whether it has addressed the harm it targeted**. This includes measuring the impact of the change on existing users, assessing how to provide proper notice of the change, and providing the proper mechanisms for enforcement.

We have published a [White Paper](#) which provides further detail on the process set out above.

Our cross-product approach to risk assessment

While the precise risks will vary by product, we consider the following overarching types of risks when considering what safeguards and rules may be needed for each product and service:

- **Encouraging harmful or dangerous behaviour:** content that either depicts particularly harmful or dangerous behaviours, or encourages users to engage in those behaviours.

- **Hateful content:** Content that promotes or condones violence against individuals or groups based on characteristics like race, ethnicity, gender identity and religion.
- **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 organisation.
- **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.

Risk assessment and the new regulatory framework

The risk assessment process outlined above demonstrates the complexity of identifying and evaluating risk and the careful, considered approach we take to developing policies. It is an approach which requires tapping into multiple areas of expertise within and beyond our company.

We look forward to seeing the guidance that Ofcom provides on risk assessments.

As Ofcom considers its approach to risk assessment, we believe the following principles will be valuable in providing industry with the necessary clarity to assess risk effectively:

Legal clarity: We welcome guidance on risk assessment that will help resolve some ambiguities in the Bill around expectations on services. The Bill requires services to carry out a "suitable and sufficient" risk assessment, and provides some indication of what could be included, but we would welcome further guidance on the desired level of detail, measurement and evaluation criteria, format of presentation, and frequency.

Practical expectations: It is also important that expectations are set so that services can practically implement them, while avoiding unforeseen impacts. For example, the

requirement in the current Bill to conduct risk assessment “before making any significant change” would benefit from further clarification as to what type of change is envisaged. According to one interpretation, this could delay our ability to make rapid changes to our services in response to an immediate threat, such as an outbreak of war and wave of disinformation.

Flexibility to accommodate differences in approach: Ofcom’s guidance can provide welcome clarity on its overarching expectations for services, but we believe that this guidance should accommodate variations in approach between services. What works for one platform may not translate across well to another platform, given the inherent differences in functionality and user base. We regularly review risk assessment processes, including in light of forthcoming regulations, as well as to account for reasonable differences in approach to also allow services to innovate in developing new methods or factors when assessing risk.

Question 4: What are your governance, accountability and decision-making structures for user and platform safety?

Is this response confidential? – ~~Y~~ / ~~N~~ / Part (delete as appropriate)

Governance and accountability

Good corporate governance is critical to our approach to content responsibility. The foundation for managing online content and conduct risks is a clear internal structure, together with accompanying processes and tools to manage the needs of Google’s diverse products and services with consistency and appropriate flexibility.

Our governance structures relating to risk include:

- Board governance: The [Audit and Compliance Committee](#) of the Alphabet Board is responsible for reviewing content-related risks. This includes issues related to privacy, safety, security, freedom of expression, and human rights. Risks related to content issues are reported to the Alphabet Audit and Compliance Committee at least annually. This ensures Board-level accountability for the safety of our users.
- Executive oversight: Senior Management has direct oversight of risks, ensures appropriate resourcing and accountability for managing risks, reviews escalations and significant risks, and reports these items to the Board as necessary.
- Program management: The responsibility for identifying, understanding, mitigating, and preventing risk of harm to users is operationalised by core functions across our Trust & Safety, Legal, Engineering, Public Policy, and Compliance functions. These teams operate horizontally across all Google

product areas and use a variety of tools to address the risk of harm related to content.

This work to identify and address the risk of harm occurs across the product life-cycle, from product development (where we embed safety into the design of our products, as we discuss below) to product launch, followed by ongoing monitoring and evaluation. Internal policies, guidance, regular training, and clear escalation paths support coordination across the structure.

For example, our Trust & Safety function is responsible for developing and enforcing product policies and, as part of Google's annual risk assessment process, our Google and YouTube Trust & Safety teams undertake a joint "content responsibility" risk review. This process involves reviewing key metrics, emerging trends and user feedback, and interviews with internal experts and executives across different subject matters and functions with a view to identifying, assessing and raising awareness of and reporting on critical user safety risks, which are then presented to the Board.

Policies and frameworks

The Alphabet Group's conduct is guided by the principle, "do the right thing". This means, among other things, following the law, acting honourably, and treating co-workers with courtesy and respect. Further specifics are set out in the [Code of Conduct](#) which is adopted by the whole Group. All Group employees, board members, extended workforce and those doing business on behalf of the Group are expected to comply with the Code, and are responsible for understanding, promoting, and implementing the Group's [guiding principles](#).

Ultimately, we see ourselves as being accountable to the public, which includes all of our users and also content owners. This is why, rather than waiting for legislation that requires us to take specific steps to minimise harm, we have acted of our own volition to protect users.

Our response to [Question 3](#) above explains how we design policies, rules and safeguards across all our products and services to protect users from the risk of harm while supporting the purpose of a product. As indicated, our policy creation process is based on risk assessment, and we consider expert input, user feedback, and regulatory guidance to help us to identify emerging harms and gaps in our existing policies.

In addition, we rely on the [UN Guiding Principles on Business and Human Rights](#) (UNGPs) as our foundational framework governing business and human rights. Google's Human Rights team regularly conducts human rights due diligence across all Google products to identify, prevent, mitigate and account for how we address any adverse human rights impacts. This due diligence includes four key steps: (1) assessing actual

and potential human rights impacts; (2) integrating and acting on the findings; (3) tracking responses; and (4) communicating about how impacts are addressed.

In addition to our internal policies and global standards like the UNGPs, we also observe multi-stakeholder and industry standards developed by entities like Global Network Initiative (GNI), Digital Trust & Safety Partnership (DTSP), the Global Internet Forum to Counter Terrorism (GIFCT), and the Tech Coalition (an alliance of global tech companies working together to combat child sexual exploitation and abuse online).

Specific structures around safety

We have a range of teams across different business functions, including Trust & Safety, Public Policy, Health, and Legal, who are dedicated to securing user safety on our services, from product development through to use by users. These teams include experts from various specialised backgrounds, including health care, child development and child psychology, who understand the risks posed by different types of content, in particular for the most vulnerable users.

Our Trust & Safety team has a specific mission to promote trust in Google and ensure user safety. The team includes analysts, policy specialists, researchers, engineers, data scientists and more, in order to develop sound, data-driven and scalable policies and standards. The Trust & Safety team consults with teams across Google to build on behalf of users globally, working to understand local context and nuance, asking hard questions and challenging the status quo, to solicit and incorporate a diverse set of perspectives. Trust & Safety guides Google in the development and presentation of trustworthy products that respect our users, and sets policy for the responsible use and access to our products and platforms that balance individual and societal rights. The Trust & Safety team draws on internal and external expertise, including academia and industry key opinion formers whose expert advice plays a role in policy development and assessment.

There are a number of other teams across Google which are dedicated to user safety, for example:

- A designated “T&S Intel” team (within our Trust & Safety team), which plays an important role in ensuring safety. For example, this team oversees processes which are intended to manage sensitive events and drive incident management protocols designed to identify, escalate, and mitigate (where possible) certain content moderation issues relating to user safety. YouTube has its own “Intelligence Desk” (noted in our response to [Question 3](#) above) which carries out an equivalent role for YouTube.

- We have various content safety steering committees, both within specific product teams and across the business more generally.

We also have a dedicated Google Safety Engineering Center in Dublin, with experts working to tackle the spread of illegal and harmful content.

Safety by design

At the product development stage, the experts noted in the section above work closely with the product teams to ensure that potential risks, as well as user safety, are well understood and taken into account at the earliest stage of product design. The product development stage includes building tools to facilitate the review, reporting and removal of abusive content (such as CSAM), and developing new and improved detection methods.

To take an example, for **YouTube**, there is a Cross Product Safety team which assesses the risk of new products and features.

On **Search**, we follow a rigorous launch evaluation process (explained [here](#)) that includes the following stages:

- First, we identify signals that can differentiate between content that should be ranked higher versus content that should not rank as highly (because it is irrelevant or otherwise low quality – as described in our [Search Quality Rater Guidelines](#)).
- Second, we implement a change (through code) to reflect these differences.
- Third, we subject this change to rigorous tests:
 - Possible changes to Search are run across a range of automated tests to ensure that they don't create obvious issues (for example, crash Search or alter far more results pages than would be expected).
 - We have these changes evaluated by our [Search Quality Raters](#) to confirm that the changes are working as expected and do not create unexpected quality "losses" where our results are altered in ways we didn't expect.
- Finally, we experiment by rolling out the change to a randomly-selected population of our users (often 1%) and study the experimental data that comes back. This process can take time, as our users need to be exposed to a sufficient number of queries where the ranking change makes a significant difference.

- Before rolling out fully, launches are reviewed and approved by a cross-functional group of experts and product leadership.
- Once all of our testing validates that the change is working as intended, we roll it out more fully. This rollout process is itself staged (often over a few hours or days) to ensure that no unexpected problems arise.
- Even after that, we may “hold back” a small fraction of traffic in order to provide a control group allowing us to continually assess the effects of the change over time – something that is particularly important where users’ behaviour may change over time.

This disciplined process helps us to avoid issues such as launching updates that create bugs and errors, as well as ensuring that the proposed change is truly more helpful. There are cases where a proposed ranking change works well for one class of queries, but quite poorly for another. Only by constantly assessing the effects of our changes can we verify that we are, in net terms, making Search better with each change.

Consistency in consideration of user safety

We tailor our policies for each product and service to strike the appropriate balance between providing access to a diversity of voices and limiting harmful content and behaviours. However, we ensure consistency in our approach to user safety through, for example:

- Top-down responsibility, and accountability at senior management and Board level, as explained above.
- Consistent principles, such as our [Code of Conduct](#) and Group [guiding principles](#), which apply across our business.
- Cross-product structures and processes, such as the specific teams noted above.

Question 5: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?

Is this response confidential? – Y / N (delete as appropriate)

Our terms of service and policies

We have universal [Terms of Service](#), which apply across a range of Google services. We also have service-specific additional terms and policies for many of our services, which are available on one page (see [here](#)). Our terms of service explain, through the use of simple and plain language, how those services work and the user’s relationship with Google. These terms of service constitute a legal agreement between Google and the user. They require the user to comply with our policies, which explain what behaviour

and content is and is not permitted on our services.

A range of supporting materials and resources, across our services, provides additional information to help further explain our policies to users. The use of separate content policies and supplementary materials can also help to ensure that our more formal terms of service do not become overly detailed. Given the pace of changes in policies related to content moderation, including every detailed update to our content policies in our terms of service would become overly burdensome for users, who would be constantly bombarded with updates.

Across our services, we aim to make our terms of service and content policies clear and easily accessible to all users and content creators. We avoid using legal jargon and, in some cases, we also use video explainers to make sure that our policies are as clear and accessible as possible. Our response to [Question 6](#) provides more detail on this.

What providers can do

From our experience, to enhance the clarity and accessibility of terms of service and public policy statements (as further explained in our response to [Question 6](#)), providers can publish terms of service and content policies that are publicly available in clear, plain language and accessible formats, and they can make publicly available change logs which show the history of changes to the terms of service and explainers of how content policies are updated.

To supplement terms of service and content policies, as noted above, we have found it helpful to users to create supporting materials, which are available in a centralised location, such as those identified in our response to [Question 6](#) below relating to YouTube. We think other providers could follow a similar approach to give users as much helpful information as possible in different formats.

Question 6: How do your terms of service or public policy statements treat illegal content? How are these terms of service maintained and how much resource is dedicated to this?

Is this response confidential? –Y/ N / Part (delete as appropriate)

How our terms of service and content policies treat illegal content

As explained above, our terms of service and content policies explain the content or behaviour that is and is not permitted on our services (including, but not limited to, illegal content).

Approach in our universal Terms of Service

Our universal [Terms of Service](#) provide an overarching framework that clarifies our right to remove content which:

(1) breaches these terms or our [service-specific additional terms or policies](#);

(2) violates applicable law; or

(3) could harm our users, third parties, or Google.

The Terms of Service list, as examples of such content, child pornography, content that facilitates human trafficking or harassment, terrorist content, and content that infringes another's [intellectual property rights](#).

Our approach on YouTube

YouTube has its own [Terms of Service](#) and [Community Guidelines](#). This is because of its unique features, including the ability for a large number of users to rapidly share and access video content, and the need to address the particular content issues that may arise on the service.

YouTube's Community Guidelines include clear statements on content that is not permitted on the platform. This includes, but is also broader than, the definition of illegal content under the Bill. There are specific policies particularly relevant to illegal content, including on:

- [Spam, deceptive practices & scams](#)
- [Child safety](#)
- [Suicide and self-harm](#)
- [Harassment and cyberbullying](#)
- [Harmful or dangerous content](#)
- [Hate speech](#)
- [Violent criminal organisations](#)
- [Sale of illegal or regulated goods or services](#)

Alongside our Community Guidelines:

- We provide a [summary](#) explanation of how our policies are updated, as well as a user friendly [video explanation](#).
- We explain [how policy violations are detected](#), including that we use a combination of people and machine learning, and how human flagging of content works.
- We tell our users how potential policy violations are considered against [exceptions](#), such as content that is educational, documentary, scientific or

artistic.

- We also explain [what action we take](#) in respect of policy violations, our “strike” system and how we age-restrict content, with links to more detailed resources.

Our approach on Search

Users of Search who enter queries are bound by our universal [Terms of Service](#). These terms do not bind the website operators whose pages are indexed (who do not, in general, enter into a contractual relationship with Google). Search does, however, apply a set of [content policies](#) under which we may delist certain content, even if not legally required to do so. For example, we have a policy on [spam](#), and our [quality guidelines](#) outline some of the illicit practices that may lead to a site being removed from the Google index or otherwise affected by enforcement action.

We also deploy more extensive policies for [search features](#) where Google curates content, which include knowledge panels, “top stories” carousels, enhancements to web listings, predictive and refinement features (such as auto-complete), and results and features spoken aloud. These policies don't apply to organic web results. For example, we do not allow the following categories of content in our search features:

- Dangerous content
- Deceptive practices
- Harassing content
- Hateful content
- Manipulated media
- Medical content
- Regulated goods
- Sexually explicit content
- Terrorist content

How our terms of service and content policies are maintained

We make changes to our content and other policies every year in response to the evolving nature of user behaviour, abuse vectors, cultural norms, sensitive content types, moderation technologies and our own product changes.

When we produce new terms of service and content policies, we take great care to ensure that they are clear to all users. This includes considering the needs of different categories of users, including children, parents and people with accessibility needs.

Google’s universal Terms of Service are regularly reviewed and updated, and we [archive and publish](#) previous versions of these terms. More recently, we have also published (on

this same page) comparisons of each version to the previous version to make it as easy as possible to see what has changed.

How we monitor the effectiveness of our terms

We track traffic to the webpage [policies.google.com](https://policies.google.com/terms-of-service) which contains our universal [Terms of Service](#). In 2021, there were 425 million visits to this page over a 90-day period.

We continually review our internal standards on the language of our policies and how we can make our policies clear and intelligible to users, in areas such as how we write, format and present our policies externally. Internal guidelines are reviewed and made available to teams within Google responsible for policies

How much resource we dedicate to this

We use both internal and external resources to design and maintain our terms of service. As explained in our response to [Question 3](#) above, numerous teams across the company - including our Trust & Safety, Public Policy and Legal teams, together with our Product teams - are involved in user safety and in the enforcement of our terms of service.

In terms of our policy change process, there would typically be several product-specific, policy and legal experts who would propose changes to the language in the relevant policy. We then have teams whose responsibility it is to implement changes, manage the translation process and oversee the process of rolling out the new policies or new wording.

Terms of service and the new regulatory framework

It is important that services are clear and transparent in their policies about what type of content is prohibited and how they treat it. Users should be provided with information that is as precise and specific as reasonably possible.

In Ofcom's approach to terms of service, we would recommend that it takes due account of the risk of inadvertently exposing sensitive information to bad faith actors, such as terrorists or hostile states. Expectations on services must reflect the risk of giving users descriptions of methods and tools used in content moderation at a level of detail that could allow bad faith actors to game platforms' systems, to the detriment of user safety.

<p>Question 7: What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms?</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>We believe it is important for services to empower users with choices and ensure they have access to transparent, fair, and effective systems that can help to protect them from both harm and censorship. We provide users with clear and accessible mechanisms and guidance on how to flag content, as described below in our response to Question 8.</p> <p>Considerations for an effective regulatory framework</p> <p>We look forward to supporting Ofcom in developing clear and workable principles and standards around reporting and complaint frameworks in the Codes of Practice. We fully support the Bill's objective of ensuring that users can access prompt and effective resolution when making reports. To ensure that the new framework works in practice and delivers on this objective, we would highlight the importance of ensuring that services can take informed action. We would also stress the value of aligning reporting between different jurisdictions and regulatory regimes where possible.</p> <p>As mentioned in our response to Question 8 below, users frequently flag content to us even though the content does not, in fact, breach any of our terms, policies, or Community Guidelines. In order for services to deal with complaints effectively, service providers should have flexibility to design user reporting systems according to the specific context or product around which they are designed to operate. For instance, reports or complaints for one product may require users to provide certain information (for example, a specific timestamp in a video or an explanation of why the content should be removed) which may not necessarily be required for reports or complaints for a different product. Flexibility would allow service providers to maintain the quality of complaints and reduce the number of bad faith or invalid complaints. This in turn would allow meaningful redress to be delivered more quickly to those users who need it.</p>
<p>Question 8: If your service has reporting or flagging mechanisms in place for illegal content, or users who post illegal content, how are these processes designed and maintained?</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>As noted in our response to Question 2, we have two distinct internal processes for the removal of content: policy violations and legal removals.</p> <p>How users report content on our services and what type of content they can flag</p> <p>As we discuss below in relation to YouTube and Search, our reporting mechanisms for policy violations are designed to allow users to immediately flag content of concern and ensure that users provide the information that we need to quickly assess the content for policy violations and to take action, where necessary. The specific mechanisms for flagging content to Google vary from service to service (as reflected in our examples further below).</p>

How we design our reporting tools to be user-friendly and accessible

We have adopted the following principles to ensure that user reporting is as easy and accessible as possible:

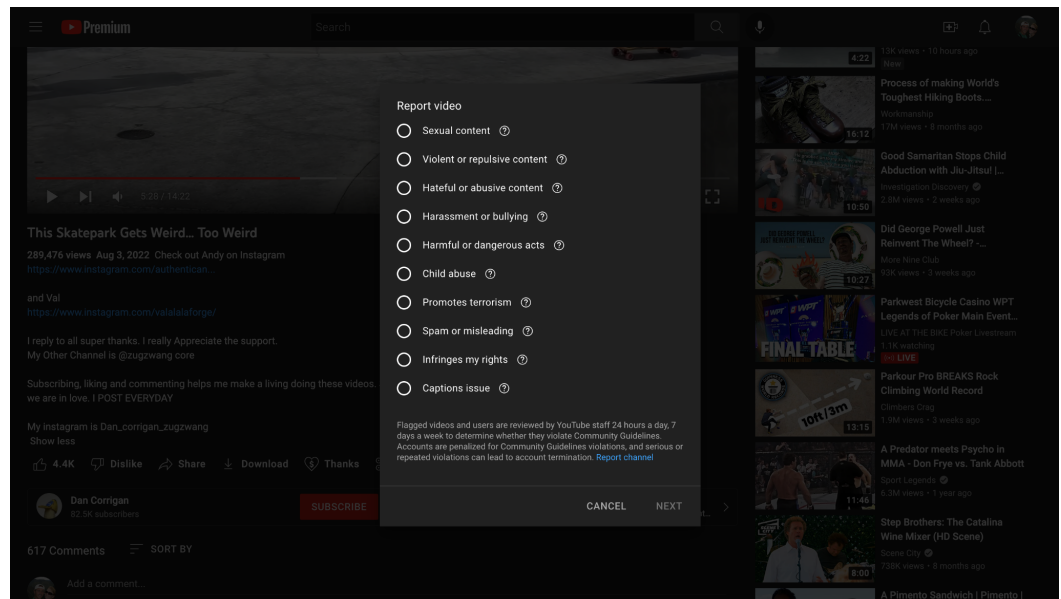
- In many of our products and services, we provide clear “buttons” for reporting content. These are key to user engagement with the complaints process.
- We provide detailed and user-friendly information for users on how to make complaints. This means that users who may not notice or understand the buttons for flagging or complaining about content are still able to guide themselves through the process.
 - For example, Google’s Help Center (available [here](#)) provides information for users on reporting, as well as information on how users can report inappropriate behavior towards children, including grooming and other forms of child sexual exploitation.
- We include accessibility features (such as providing explanations on how to make complaints in a read-aloud format in different languages).
- We provide information for users to understand the process once they have flagged content.
 - For example, on YouTube, we have produced a video on [The Life of a Flag](#) to help users understand what happens to content they have flagged.

Our approach on YouTube

On YouTube, we allow users to select the following categorisation boxes (which include, but are also broader than, the definition of illegal content under the Bill) when reporting content:

- Sexual content
- Violent or repulsive content
- Hateful or abusive content
- Harassment or bullying
- Harmful or dangerous acts
- Child abuse
- Promotes terrorism
- Spam or misleading
- Infringes my rights
- Captions issue

These categorisations appear in a pop-up box when a user clicks on the “report” button, which is available for the relevant content (for example, videos or comments).



There are various ways in which reports can be made, including:

- Flagging a video that contains inappropriate content;
- Flagging comments which are spam or abusive;
- Filing an abuse report where there are multiple videos and comments shared by the same user, or a user’s entire account (channel), that warrants reporting; and
- Reporting autocomplete and search predictions.

It is not necessary for a user to create an account in order to use these reporting mechanisms.

For legal removals, such as defamation concerns, users can also use a [service-specific webform](#). Users are able to select the reason for requesting removal from a list of topics, including harassment, copyright or privacy, among others. Our legal removals team, comprising trained experts, reviews the report and determines whether to remove the content in accordance with applicable laws.

Our approach on Search

For Search, we have carefully developed a set of [content policies](#). This page also provides users with information on how to report specific types of content that violates those policies (such as personal information or spam). We have [reporting tools](#) for requesting the removal of personal information (such as NCEI or doxing content) from Search. We also have in-product reporting tools for many Search features, such as

auto-complete, and similar feedback mechanisms for other Search features, such as [knowledge panels](#) and [featured snippets](#).

In addition, we have created an intuitive [troubleshooter](#) for any user who wants to make a legal removal request on any of our services, including Search. This troubleshooter allows users to report content as being illegal, including because it is a scam, malware, copyright infringement, defamatory statement, or content designated to be unlawful by a court. In the case of child abuse imagery, we also show users messages directing them to independent specialist organisations, as described further in our response to [Question 12](#) below.

As mentioned in our response to [Question 2](#) above, because we do not host content on the web, we cannot entirely remove content; this can only be done by webmasters. However, we do provide [information](#) to help users understand how to request removals from hosting websites and webmasters. We also provide notices in our [webmaster console](#), so that those hosting content can be notified and take the appropriate action, whether to contest the delisting, update their web page, or remove it altogether. Individuals can also notify us of outdated content with the [Outdated Content Removal tool](#).

Case study: NCEI on Google Search

We aim to ease the process of making removal requests as much as possible, particularly in relation to the most sensitive content. To provide an example, we have taken the following steps in relation to Google Search and NCEI (also known as “revenge porn”):

- First, we have built [reporting tools](#), which allow victims (or their authorised representatives) to report content for review under these policies (as shown in the screenshot below):

Request to remove your personal information on Google

Use the options below, to contact Google about a personal information removal.

What do you want to do?	Remove information you see in Google Search	
Let us know where you saw the information you want to have removed.		
The information I want removed is:	In Google's search results and on a website	
Have you contacted the site's website owner?		
	No, I prefer not to.	
I want to remove	Nude or sexually explicit items	
	A nude, sexual or intimate picture or video	
Are you (or someone you are authorized to represent) in the images or videos and are you nude or are they otherwise sexually explicit? Yes		
Have you ever consented to the distribution of the images or videos? No		

- Second, our reporting tool allows people to upload multiple URLs so that any content a person identifies in search results can be reported through a single form:

URL(s) of the webpage(s) that show the content ([Learn how to find the URL](#)) *

Please enter one URL per line (Max 1000 lines)

- Third, we provide an option for individuals to request that Search filters explicit results for Search queries similar to the one included in the removal request (shown in the screenshot below). For example, if a user's removal request is related to the query [Joe Bloggs leaked nudes], and that request is approved, then we may filter explicit results from that query going forward. The goal here is to mitigate the need for users to continually re-input removal requests.

I would like Google Search to filter explicit results for similar searches in the future.
Note: While Google will do its best to filter explicit results, we cannot guarantee that we'll catch everything.

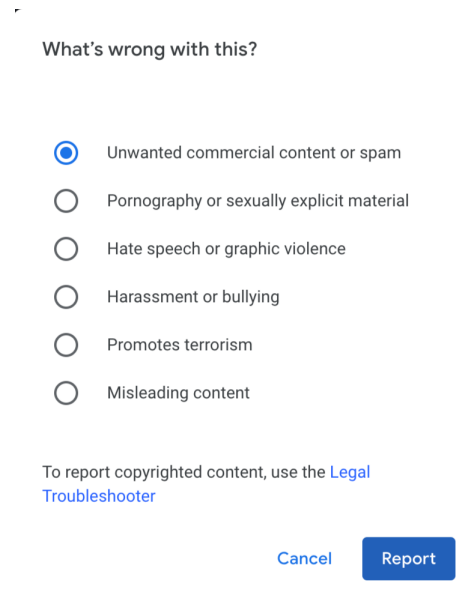
- Fourth, for image URLs that are reported via the NCEI reporting tool, and found to be violative and subsequently de-listed, we have systems in place to detect and remove copies of this content from Search. While Google makes best efforts

to stop this content from appearing, images can easily be modified and therefore evade detection via current hash-matching technology. As a result, these “de-duplication” protections may not detect all manipulated but visually similar “near-duplicates.”

Our approach on other Google services

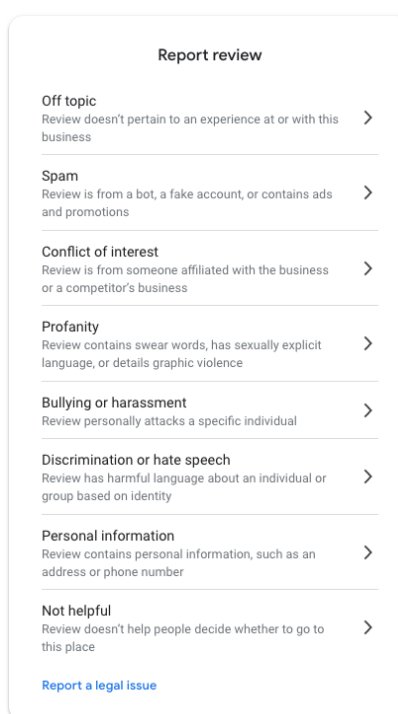
There are similar mechanisms that allow users to report content to us across our other services, such as Google Photos and Google Maps. These mechanisms necessarily vary depending on the policies for that specific service and the nature of the content on the service.

For example, on **Photos**, users can report content to us using the “Report Abuse” link in-app (as explained [here](#) and shown below):



The screenshot shows a dialog box titled "What's wrong with this?". It contains a list of radio button options: "Unwanted commercial content or spam" (selected), "Pornography or sexually explicit material", "Hate speech or graphic violence", "Harassment or bullying", "Promotes terrorism", and "Misleading content". Below the list, there is a link: "To report copyrighted content, use the [Legal Troubleshooter](#)". At the bottom right, there are two buttons: "Cancel" and "Report".

On **Maps**, users can report or flag content (as explained [here](#)) - including reviews, photos, videos, and questions or answers - on various grounds, including for example, because it is spam, contains profanity or amounts to bullying or harassment:



Reports from third parties, including Trusted Flaggers

While anyone can report content to us, we design our reporting mechanisms so that different groups of people can report content to us:

- **Courts and government agencies** around the world regularly request that we remove information from Google products (including YouTube and Search). We review these requests closely to determine if content should be removed because it violates a law or our product policies. Our teams assign each request a category, such as hate speech, obscenity, or defamation. Between 1 July 2021 and 31 December 2021, we received 28,913 government requests to remove 385,396 items of content from our platforms. Of these 28,913 requests:
 - 6,694 were due to copyright concerns;
 - 4,189 were due to national security concerns;
 - 2,238 were due to concerns regarding regulated goods and services;
 - 2,323 were due to concerns that the content was defamatory;
 - 5,989 were due to privacy and security concerns; and
 - 7,480 were for other reasons.
- We also have **Trusted Flagger** programmes across a number of our services.
 - For example, through our Trusted Flagger programme for **YouTube**, we have around 300 government partners and NGOs with expertise in areas such as child safety, hate speech, and violent extremism, who receive training in identifying content that violates our Community Guidelines.

Their flags are prioritised for review by our policy and enforcement specialists, who review and take action on the content where appropriate.

- There are other ways that certain **third parties** can report content. For example, in addition to our own extensive proactive monitoring, Google acts on URL lists provided by child safety organisations for removal purposes on Search, but also in order to add to the body of known CSAM that can be detected through hash matching (discussed below in our response to [Question 12](#)). This includes organisations such as the Internet Watch Foundation (IWF) and NCMEC.

Effectiveness of reporting and flagging mechanisms

While they are an important way of identifying problematic content in addition to our own classifiers, user flags are not typically very accurate at identifying illegal or harmful content. For example, on YouTube, users often flag videos to express dislike for the video rather than for the purposes of reporting content that is illegal or a policy violation.

That said, we believe we have an effective mechanism for removing content on YouTube once it is flagged to us. For example, of the 49,256 videos that were removed on YouTube (after having been uploaded in the UK) in Q1 2022, only **4,330** were appealed (i.e. less than 9%) and just 905 reinstated (i.e. less than 2%). This scale of accurate removal is only possible due to machine learning classifiers, with 41,834 (almost 85%) of those videos removed in Q1 2022 first detected by automated flagging and 25,297 of UK videos removed before they received 10 views.

Notice formalities for illegal content

We have found it helpful to have a standardised and substantiated process for formal user requests to remove illegal content. Notice formalities help review teams process information more efficiently and responsibly, as well as protect against abuse by fraudulent or bad faith actors. By notice formalities, we mean, for example:

- clear identification of the content at issue by URL, video timestamp, or other unique identifier in a tangible and usable format;
- identification of the law and basis of the legal claim;
- clear identification of the sender of notice where the nature of the rights asserted requires identification of the rightsholder; and
- attestation to the good faith and validity of the claim.

	<p>Other means to dissuade bad actors from submitting fraudulent or false claims should be considered. This is a known problem that can significantly slow down the review of other, valid notices of illegal content.</p>
<p>Question 9: If your service has a <i>complaints</i> mechanism in place, how are these processes designed and maintained?</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>We have explained in our response to Question 8 above how users can report content to us. In response to this question, we focus on our mechanisms for responding to complaints or “appeals” to decisions we make to restrict or remove content that has been generated, shared or uploaded on YouTube by users, and the limited cases in which we allow appeals against the delisting of content on Search.</p> <p>On YouTube:</p> <ul style="list-style-type: none"> ● When a creator’s video is removed due to a policy violation, we provide a link with simple steps to appeal the decision. ● If a creator chooses to submit an appeal, that appeal is reviewed by a member of our Trust & Safety team. ● We keep detailed records, including data on complaints, appeals and reinstatements, and we think it is important to be transparent about this information, which we publish here. This report includes data on video, channel, and comment removals; appeals and reinstatements; and human and machine flagging. ● Our “three strikes” approach to moderation on YouTube enables us to balance our aims of keeping users safe online while also preserving freedom of expression for our creators, as well as allowing us to educate creators about our Community Guidelines before removing their videos or channels. <ul style="list-style-type: none"> ○ We understand mistakes happen and creators don’t mean to violate our policies - that’s why the first violation is typically only met with a warning. ○ For a subsequent violation, we issue a “strike” against the channel. Strikes also come with upload freezes, meaning that creators who receive a strike are barred from uploading to the platform for one week on their first strike and two weeks on their second. Approximately 94% of people who receive a first strike never receive a second one. ○ If a creator receives three strikes within 90 days, their channel, and therefore all of its content, is removed from YouTube. ● Users who repeatedly or maliciously report content or otherwise misuse our complaints mechanisms may have their accounts suspended and be prohibited

from using them.

On **Search**, we also provide appeals processes for removals on some content types, such as [copyright](#), [NCEI](#), and [counterfeit goods](#), as well as for removals under some specific pieces of law, such as EU data protection law. We are committed to providing necessary and proportionate mechanisms for appeals and/or requests for reconsideration, in areas of identified need. That said, appeals alone cannot serve as a backstop for overly broad removals or demotions at scale, taken in response to general monitoring provisions, takedown-staydown requirements, and/or other minimisation obligations that leave no other room for compliance aside from sweeping, automated filters.

In addition, search engines generally lack a direct, contractual relationship with content creators. Often, the only information Google possesses about a given website is what is available to anyone with a web browser. In this regard, appeals as commonly understood in the context of social media services do not function the same within Google Search. Unlike hosted platforms, where users sign in and post content, webmasters do not sign in and “post” their URLs. Instead, Google Search - like all search engines - crawls and indexes the open web. This makes offering appeals more challenging on a search engine than on a hosted service. Where feasible, we provide notification of actions taken in Search via our webmaster-facing [Google Search Console](#) if the affected site has registered.

Separately, allowing for appeals in relation to de-prioritisation or removal of content from search results would be extremely difficult from an operational perspective. A few issues complicate the feasibility, ability, and even desirability of complaints for demotions:

- **First, ranking is dynamic and technically complex:** On Search, it is difficult, if not impossible, to give appeals to all webmasters whose content is "demoted", given that ranking is not static; for example, results may rank differently as new content is published on the open web. Also, as Google launches improvements to its ranking algorithms that improve the user experience, these necessarily result in some sites ranking less well, all other things being equal. Many of these improvements happen deep in our systems and at a very high level of generality; they do not target the ranking of individual websites. For instance, if we launch an improvement to our ability to understand synonyms, that might result in a change to our ranking on certain queries. To allow every site incidentally affected by this to impede the service's improved understanding of language would hamstring our ability to innovate for our users.
- **Second, search ranking is different from ranking on a newsfeed:** Something that is ranked highly for one query will not necessarily rank highly for another query. This is not a "demotion" per se, but the nature of Search ranking

operating in response to queries, rather than to populate a "feed".

- **Third, complaint mechanisms for demotions are unfeasible on a practical level:** Offering appeals for algorithmic ranking decisions does not work for a product like Search for the reasons stated above. Moreover, user redress mechanisms for such decisions would be rife for abuse by spammers, scammers, and others who would litigate ranking decisions for illegitimate purposes. Scale is an issue here that needs to be contemplated when debating these types of obligations. Such mechanisms would be overwhelmed by those who would want their content to rank more favourably, for a multitude of reasons that do not relate to relevance for users.

We offer a reconsideration [option](#) for site operators whose content has been deprioritised as a result of a [manual action](#) under our [Webmaster Guidelines](#) (which prohibit technical manipulation of our ranking, also known as "webspam"). We do not, however, offer a general complaint mechanism for site operators whose content is not ranking as well as they might like.

Question 10: What action does your service take in response to reports or complaints?

Is this response confidential? – Y / N (delete as appropriate)

With reference to [Question 2](#) above, we refer again to our two distinct, internal processes for the removal of content - policy violations and legal removals - because the process we follow on user reports depends on the exact reporting channel used by the user and the product in question, as explained below.

Our approach on YouTube

When users flag videos on YouTube, trained teams evaluate the videos before taking action in order to ensure that they actually violate our policies and to protect content that has an educational, documentary, scientific, or artistic purpose. Reviewers evaluate flagged videos against all of our Community Guidelines and policies, regardless of why they were originally flagged. The teams, operating 24/7, carefully evaluate flags. They remove content that violates our terms, age-restrict content that may not be appropriate for all audiences, and leave content live when it does not violate our guidelines.

We have provided further information on the process explained above in our video on [The Life of a Flag](#) to help users understand what happens to content they have flagged.

Our [YouTube Community Guidelines Enforcement report](#) provides the total number of user flags and breaks down the type of flags we receive according to different categories of violation of our policies, by quarter. For the most recent quarter (Q2 2022), 21,246,972 videos were flagged by users globally. To put this into context, 4,496,933

videos were removed in the same period; of these, only 256,109 were first identified by individual users (trusted flaggers and government requests are reported separately). The highest proportion of flags were received in relation to our policies on: Spam or misleading content (26.7%); hateful or abusive content (21.1%); and sexual content (19.1%).

Reviewing flags requires careful and nuanced judgments by human reviewers. The context of content can be very important. The same piece of content with a different context can indicate very different intent from the user. For example, there is a clear difference between content intended to shock or disturb a viewer, potentially causing them harm, and content which documents real-world events where there is a public interest in keeping that content available. Similarly, certain terms/phrases often used to direct hate towards particular communities may be perfectly legitimate when used by members of that community. In addition, as explained above in response to [Question 8](#), user flags can have very low accuracy rates, especially flags made in bad faith or based on a dislike of a particular video.

Our approach on Search

Because Google Search does not host user generated content in the same way as our user-to-user services, content removal on Search is largely (although not exclusively - for example, we allow users the possibility of removing their personal information from Google Search) focused on legal removals rather than policy violations.

Our [transparency reporting](#) provides further detail on the action we take in relation to content removal requests we receive for Search. For example, for the period of July to December 2021, for all content removed in response to a UK government request or court order, 59.5% of that content was removed through a legal removal, whereas only 4.3% was removed for a policy violation. This is largely because, as explained in our response to Question 1, user expectations are different for search services than they are for user-to-user services, and so content policies for Search are not as extensive as on most user-to-user services.

While the process we adopt in response to a report varies depending on factors such as the nature of the complaint, in general, the process or actions we take are as follows:

- We receive reports through our webforms, which are then routed to the right teams within our Trust & Safety team (depending on the basis of the complaint).
- For certain reports, we use automated tools to perform a first-cut analysis (for example, Digital Millennium Copyright Act (DMCA) notifications for copyright infringement are first assessed by an automated tool). For others, an individual reviews each and every report.

- The reviewer then evaluates the report under the appropriate policy or law, and escalates to the appropriate team internally.
- The reporting user receives a response via email from the relevant team.

The rates at which we approve removal requests vary from reason to reason. For example, we remove in excess of 90% of the content we are asked to remove on copyright grounds (for which the analysis is generally relatively straightforward), and we remove about 50% of the content we are asked to remove under the “right to be forgotten” ground (for which the analysis is not as straightforward).

Notifications to third parties

When we become aware of certain illegal activity occurring on our platform, we have processes in place to proactively refer imminent threats to law enforcement. These processes enable our various internal product teams and external sources to escalate potential threats and criminal activity if they see it on our platforms. Google's dedicated analysts then assess escalated threats, and, where appropriate, make proactive referrals to law enforcement.

We apply special tools and processes to reports of CSAM, including the following:

- Google creates a report called a **CyberTip**, which involves reporting the violative content and identifying information about the user that uploaded the material (if available) and sending it directly to the NCMEC CyberTipline (a specialist hotline to receive reports of sexual exploitation of children). NCMEC will report the matter directly to the relevant law enforcement agencies around the world, using a dedicated virtual private network. In the case of the UK, it is provided to the National Crime Agency's Child Exploitation and Online Protection unit.
- We remove the violative content from where it surfaced. Google may hash the content, so that it can be used to detect matching content across the platform (we explain hash matching in more detail in our response to [Question 12](#) below).
- On Search, when we detect URLs that contain CSAM, we will de-index the URL from surfacing on Search and make a report to NCMEC.
- We have also launched a [Transparency Report](#) on Google's efforts to combat CSAM, where we detail the number of reports we have made to NCMEC. This report also provides data around how we detect and remove CSAM results from Google Search and how many accounts are disabled for CSAM violations across our services.

Question 11: Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?

Is this response confidential? – Y / N (delete as appropriate)

We agree that there can always be improvements made in the development and enforcement of content moderation.

Across Search and YouTube, we are constantly working to improve and evolve, introducing new policy changes, hiring new people dedicated to safety policy, and continuing to invest in technology to help us tackle illegal and harmful content at scale.

We are also mindful of the need to avoid unduly restricting user activity. With that in mind, we welcome the Bill's protections for freedom of expression, as well as Ofcom's interest in the subject and in ensuring the Bill's protections for freedom of expression are given practical consideration.

Considerations for the regulatory framework around content moderation

As Ofcom considers how best to drive improvements in the protection of user safety and users' rights, we would like to highlight the following considerations.

Expanding the use of what works: As explained in our response to [Question 12](#) below, we use various tools to protect users from harmful content. We also make many of our tools, including Content Safety API and CSAI Match as part of our Child Safety toolkit, available to other companies to strengthen the ecosystem. Our Child Safety toolkit helps partners to classify over 4 billion pieces of content per month. We would be very keen to work with Ofcom and others in considering how we can expand the use of these tools and other effective technology we use to tackle such content.

Instilling safeguards for users' rights: Our experience of the opportunities and pitfalls of using technological tools underpins our concern that the regulatory framework could, without further guidance from Ofcom, incentivise the use of these tools without appropriate safeguards for users' rights. Our experience is that these tools can struggle to identify illegal content accurately where doing so requires an understanding of contextual nuance, due to the nature of that content (for example, in relation to hate speech). We would welcome the opportunity to work with Ofcom on how providers of online services can most effectively identify illegal content in these circumstances.

Preserving flexibility for innovation: It will be important for the regime to preserve the flexibility of providers to continually innovate and stay ahead of bad actors. In a fast-moving threat landscape, the most effective regulatory approach is to hold providers to account for the impact of the steps they are taking, rather than to prescribe the use of specific tools that can quickly become out of date. In particular, there is a real possibility that any over-use of Ofcom's powers to direct the use of particular technology could signpost to bad actors which tools are being used across the industry, enabling

them to focus their efforts on circumventing those tools.

Maintaining protections in relation to journalistic content: We endeavour to prevent users from inadvertently coming across content that may cause them harm. For example, we add warning labels to YouTube videos to alert users to graphic content in the video, which mitigates the harm of some users inadvertently viewing that content without unduly restricting other users from intentionally viewing the same content. An amendment to the Bill made during the Report Stage prevents providers from “taking action” in relation to content published by a user which is a recognised news publisher. We have conveyed our concerns that such a temporary “must carry” obligation is ripe for abuse. In light of the broad definition of “recognised news publisher”, this obligation risks a significant volume of harmful content being accessible while service providers engage in a time-intensive “ping-pong” appeals process with the content provider. This broad definition also creates a risk that service providers may be required to retain harmful content from extremist sites or state-backed or state-owned news providers. Recent events, notably the war in Ukraine and COVID-19 pandemic, have reaffirmed how vital it is to redirect users away from low quality, unauthoritative, and untrusted sources in a time of emergency and isolation. We would welcome further discussion on the process for determining a “recognised news publisher.” We would also welcome clarification from Ofcom in the Codes of Practice that this would not prevent us from, for example, adding a warning label to graphic images or videos published by a recognised news publisher while an appeal is pending.

Question 12: What automated moderation systems do you have in place around illegal content?

Is this response confidential? – Y / N (delete as appropriate)

We have long invested in the most effective automated systems for protecting users from harmful content and have developed effective automated detection tools, which we have opened up to the wider industry.

We focus this response on the automated systems we have in place for CSAM, specifically. This is the area in which our solutions are the most advanced because the legal framework is clear (CSAM is nearly universally illegal) and the binary determination of CSAM or not-CSAM is usually relatively clear and less context-dependent. Machines also can help to flag hate speech and other violative content, but these categories are highly dependent on context and require human review to make nuanced and accurate decisions. We agree with the conclusions of the US Federal Trade Commission (FTC) [report](#), published in June this year, which warned against over-reliance on automated technology to proactively monitor for illegal and harmful content, because of concerns about blocking legitimate content and the impact on free speech. The FTC’s report stressed:

“it is crucial to understand that these tools remain largely rudimentary, have substantial limitations, and may never be appropriate in some cases as an

alternative to human judgment. Their use — both now and in the future — raises a host of persistent legal and policy concerns. The key conclusion of this report is thus that governments, platforms, and others must exercise great caution in either mandating the use of, or over-relying on, these tools even for the important purpose of reducing harms”.

Our response to [Question 13](#) below explains more about the safeguards we have in place for users’ freedom of expression and privacy in circumstances where we think it is appropriate to use automated tools for content moderation.

Automated systems for CSAM

We note the following examples below of our automated moderation systems on YouTube and Search for CSAM:

- **Hash matching technology:** we have been using hash matching technology since 2008 to detect known CSAM and we have continued to improve and adapt this technology. This technology allows for automated review with a high level of precision and accuracy, meaning that flagged content can be removed straight away. For example, where a hash match confirms that virtually identical content has been removed elsewhere, the matching content can be removed immediately using only an automated tool. We make this technology available to NCMEC in the Hash Matching API. The Hash Matching API identifies visually similar files for a given image, which has enabled NCMEC to triage the high volume of images they receive through their CyberTipline more efficiently.
- **Machine learning classifiers and the Content Safety API:** We have invested heavily in developing machine learning tools that allow us to detect new, not previously seen CSAM, at scale. The tool helps to identify content that is likely to be abusive, which is then prioritised for human review. In 2018, we started making this technology available to others in the form of the [Content Safety API](#), to support their efforts in detecting new CSAM. This enables our trusted partners, including industry and NGOs, to find and report abuse much more quickly.
- **CSAI Match:** This tool was specifically developed by YouTube engineers, building on our proprietary video technology, to support the detection of known CSAM in video format. It uses hash matching to identify re-uploads of previously identified child sexual abuse in videos.⁵ The Tech Coalition’s 2021 [Transparency Report](#) (published in August 2022) indicates that one third of member

⁵ See YouTube CSAI Match, <https://www.youtube.com/csai-match/>; using AI to help organisations detect and report child sexual abuse material online.

companies make use of YouTube's CSAI Match service. This is the highest adoption rate for any video matching technology by Tech Coalition members for CSAM detection.

Effectiveness of automated systems for CSAM

In 2021, we launched a [Transparency Report](#) to bring more visibility to the impact of our efforts to fight online CSAM. This report shows that:

- In 2021, there were 596,710 URLs in H1 and 580,380 URLs in H2 that were de-indexed from Search for violating our policies in relation to CSAM. This is a total of 1,177,090 URLs de-indexed in 2021 alone.
- In the past year, we made almost 500,000 reports to NCMEC containing 6.7 million pieces of content, including images, videos, URL links and/or text soliciting CSAM. (A single piece of content may be identified in more than one account or on more than one occasion, so this metric may include pieces of content reported more than once).
- NCMEC [reports](#) that its use of the Hash Matching API developed by Google means that it has been able to tag more than 26 million images. Further, NCMEC reports ground-breaking improvement in their reviewer teams' well-being by reducing the need to look at the same images over and over again.

More information on the technology that we use to detect CSAM, and our wider efforts to combat child sexual abuse, can be found [here](#).

Automated systems on YouTube

On **YouTube**, the majority of takedowns are made under our Community Guidelines (i.e. policy violations, rather than legal removals), which apply globally but which substantially cover most types of illegal content (for example, CSAM, violent extremism and hate speech).

We have developed a number of classifiers aimed at pro-actively detecting these types of content. According to our last reported figures from Q2 2022, over 93% of video removals were initially flagged by our automated systems, with almost one third removed before they had a single viewing and over two thirds removed before they had more than 10 viewings. However, as our response to [Question 13](#) explains, careful human review is necessary to assess much of the content that is first flagged by automated systems, to avoid the large-scale removal of legitimate content.

We have special measures in place for possible terrorist content that would violate our policies. For example, as part of YouTube’s membership and leadership of the GIFCT, we use technology to prevent re-uploads of known terrorist content before that content is available to the public. In 2016, we created a hash-sharing database with industry partners where we share hashes (or “digital fingerprints”) of terrorist content to stop its spread. The shared database currently contains over 320,000 unique hashes, including both videos and images. We should note, however, that the hashes within this database are not labelled by legality or illegality, either under UK or any other law. Hashes are added by member companies on the basis of policy violations and only for entities designated by the United Nations or perpetrator-produced content following a real-world violent extremist event. The hash-sharing database is, therefore, not able to be used as a tool for coordinated takedowns of illegal material under UK law, as it includes hashes of policy violative content which may or may not be illegal in different jurisdictions according to different countries’ standards and processes for designations.

The best method for assessing the effectiveness of our content moderation efforts on YouTube is our VVRs, which we describe in our response to [Question 2](#).

Automated systems on Search

On **Search**, we use CSAM detection technology and reporting mechanisms, as described above and in our response to [Question 10](#). We also work to deter searches for this type of content. When we detect that a query in Google Search may be associated with child sexual abuse, we turn on additional protections, including disabling autocomplete and providing non-explicit results (for example, results excluding pornographic content). We also apply our ranking protections, which surface high-quality links, such as those that refer people to NGO resources.

In some jurisdictions, including the UK, users seeking this content are shown a deterrence message in an information box at the top of the Search results page. This deterrence message informs the user that CSAM is illegal, provides information on how to report the content to the IWF and offers those who may be concerned with their own CSAM-seeking behaviour with information about [Stop It Now](#), a campaign run by the Lucy Faithful Foundation.

Question 13: How do you use human moderators to identify and assess illegal content?

Is this response confidential? – Y / N / Part (delete as appropriate)

Our automated tools are highly effective for what they are designed for and they are responsible for identifying the vast majority of the content that we remove. However, it would be impossible for us to operate effective content moderation processes without investing a large amount in human reviewers. This is because automated systems cannot act as effective content moderation tools where content requires a more nuanced

determination. Over-reliance on machine learning technology and other automated tools to monitor services and identify illegal content poses real risks for freedom of expression because it can result in the removal of legitimate content and even content that has significant public interest importance.

As an example of what can happen when automated removal is prioritised over careful human review, in Q2 2020, as COVID-19 lockdowns meant that fewer human content moderators were able to work, YouTube depended more heavily on automated technology to remove content violating our policies. The number of appeals by users of content removal decisions doubled as compared to Q1: 50% of appeals resulted in reinstatement in Q2, compared with less than 25% in Q1.

We have teams around the world who review flagged content for policy violations on **YouTube**. We also have human review of legal removals, for both **YouTube** and **Search**, because they require an assessment of whether the content contravenes specific local law, which cannot be determined by an algorithm alone.

Human moderators on YouTube

At Google, we have nearly 22,000 people dedicated to monitoring content on our platforms. Human reviewers work around the world, 24/7, speak 60 different languages and are highly trained.

Our human moderators decide whether to:

- remove content - where it violates our Community Guidelines;
- restrict access to the content (for example, based on age where the content is not appropriate for all audiences); or
- leave the content live when their judgement is that it doesn't violate our guidelines.

Our moderators receive regular training. This training is updated when new policies are introduced or new abuses of our services are identified.

We have a rigorous quality assurance mechanism for moderators where we assess moderation decisions for accuracy (whether the Community Guidelines are correctly interpreted) and consistency. We draw samples regularly from the work done by our moderators, which are then re-reviewed by our specialised teams of quality analysts.

We also have "calibration sessions" which include the moderators, quality analysts and expert teams on our specific policies. Where it is not clear to these teams whether

flagged content violates our policies or not, they raise the report to our Trust & Safety team for further review.

Support for and safeguarding of moderators

Human moderation requires extreme care in terms of how reviewers cope with reviewing graphic and sensitive content. While most content moderation is not violent or graphic, some of the material these moderators review can be disturbing and upsetting. Some moderators choose 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. We are committed to ensuring that moderators have the highest standard of support and we have invested significantly in these teams.

Our support for these moderators includes the following:

- 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 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 (available as both opt-in and scheduled).
- Providing access to quiet rooms and community spaces, which are required at all sites.

Use of third parties to support content moderation efforts

Sometimes, where required, we work with third-party 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. In order to ensure that our guidelines and Supplier Code of Conduct are respected by these providers, we undertake regular site visits and audits.

These visits include one-on-one conversations and focus groups with reviewers so that Google can receive direct and confidential feedback from the individuals. All the third parties we work with provide their employees with grievance reporting and redressal

mechanisms, as well as 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.

Google research on content moderation

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 other products choose to review images in grayscale and keep the grayscale option turned on. We’re committed to rolling out this option more broadly.

External (user) notification of content potentially in violation

We encourage members of the public, civil society groups, and authorities to alert us to content they believe may be in violation of policies and/or illegal, so that we can assess it and, where necessary, remove it from our services. For example, we have a network of over 180 academics, government partners – including the UK’s Metropolitan Police – and hate speech NGOs through our YouTube Trusted Flagger programme. Participants in the Trusted Flagger programme receive training in enforcing YouTube’s Community Guidelines. Because of this training and these partners’ expertise in identifying hate speech, when they flag potential hate speech content to us, we prioritise it for review.

Question 14: How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services?

Is this response confidential? – ¥/ N (delete as appropriate)

Google’s services (including **YouTube** and **Search**) can use a range of tools and approaches to remove access to the most harmful content and limit the visibility and prominence of less authoritative borderline content, which does not breach our policies but comes close to doing so.

For example, on **Search**, we have designed our ranking systems to prioritise content that is most helpful in response to a user’s query. To do this, the systems identify signals - such as where words in a search appear on web pages, or how pages link to one another on the web - that can help determine which content demonstrates expertise, authoritativeness, and trustworthiness.

To evaluate whether our systems do, in fact, provide information that people searching find relevant and reliable as intended, we solicit feedback on proposed improvements from Search Quality Raters who use our [Search Quality Rater Guidelines](#). We have over 14,000 Search Quality Raters from around the globe who collectively perform millions of sample searches and rate the quality of the results according to the signals we previously established. For example, if a website or page has a harmful purpose or is designed to deceive people about its true purpose, it will immediately be rated the “Lowest Quality” on the “Page Quality” rating scale. This includes websites or pages that are harmful to people or society, untrustworthy, or spammy, as specified in the Guidelines. Raters are instructed to follow the standards outlined in Section 7.0 of the Guidelines which defines what is considered harmful.

It is important to note that no single rating - or single rater - directly impacts how a given site ranks in Search. Instead, ratings are a data point that, when taken in aggregate, helps us measure how well our systems are working to deliver content that is aligned with how people - around the world - evaluate information.

On **YouTube**, we use a “three strikes” policy, as explained in our response to [Question 9](#). We also have tools in place to prevent those users from setting up new channels and accounts.

Other sanctions on YouTube include **demonetising content** on a channel, such that the channel can no longer earn revenue through videos shared on the service, or through the application of **warning signs** which indicate to potential viewers that, for example, the content may be graphic or offensive.

We put in place safeguards by ensuring that, when we do apply sanctions and restrictions, content creators are notified and given the opportunity to appeal our decisions, and challenge unwarranted sanctions and restrictions in the same way as content removals, as described in the response to [Question 9](#) above.

We explain in our response to [Question 17](#) below the approach we take to applying sanctions and restrictions around access to our services and the content on it.

Question 15: In what instances is illegal content removed from your service?

Is this response confidential? – Y / N (delete as appropriate)

Google’s approach

Illegal content is removed from our services where:

- we are able to find it, for example, using one of the tools that we deploy;

- we are able to verify that it violates one of our policies (or we are required to remove it pursuant to local law, if it is a legal removal); and
- where we have the power to remove (or, more accurately in the case of Search, delist or deprioritise) it.

YouTube

We need sufficient evidence that the content on YouTube is illegal before we take it down, as this reduces the risk of us inadvertently removing legal and legitimate content. In terms of verifying a policy violation on YouTube, in order to mitigate the risk of over-removal, we set a threshold of *requiring knowledge* that content violates our policies before removing content.

We note that, as a result of an amendment, the relevant standard in the Bill is whether providers have “*reasonable grounds to infer [our emphasis]* that content is illegal based on all relevant information that is reasonably available”.

We would welcome clarity from Ofcom in the Codes of Practice, and the guidance it will produce on illegal content judgements, on how this standard is to be applied, particularly in respect of content moderation by automated technology and in relation to offences where the intent of the content generator is relevant. This amendment suggests that platforms should apply a lower standard when assessing whether content is illegal than that applied by a court or regulator, for example. We have real concerns that a lowering of the threshold to be applied in determining that content is illegal could risk the over-removal of legitimate content without sufficient evidence that the content amounts to an offence on- or offline.

We would urge Ofcom to work closely with providers on what is technologically workable, before issuing the relevant guidance on illegal content judgements.

Search

In terms of Search, where we are notified of illegal content, we are able to delist it from our results (we cannot remove content from the web itself, as explained in our response to [Question 2](#)). Given the ramifications of delisting on users’ rights of free speech and access to information, we only delist URLs where we have knowledge that the content is unlawful through a legal removal request, to ensure accountability to Search users for sites whose pages have been delisted.

We rely on certain tools for delisting (as explained in our response to [Question 12](#)); it would be impossible to search all the content on the web and determine whether it is illegal or not.

Beyond legal removals, we also remove a limited set of information from Search in respect of policy violations, mostly focused on highly personal content appearing on the open web. Examples of this content include financial or medical information, government-issued IDs, and intimate imagery published without consent.

The tools that we use are described in more detail in our answers to [Questions 12](#) and [13](#), which explain the way in which we combine innovative automated technology and human review to ensure user safety.

Question 16: Do you use other tools to reduce the visibility and impact of illegal content?

Is this response confidential? – Y / N (delete as appropriate)

As set out in our response to [Question 15](#), if we are aware of content that we have a legal obligation to remove, we disable access to it rather than simply reducing its visibility.

Reducing the visibility of borderline content on YouTube

In relation to **YouTube**, we take steps to reduce the visibility of so-called borderline content, which, although not illegal, is content that comes close to - but does not quite cross the line of - violating our Community Guidelines.

Our goal is for recommendations and for search results on YouTube to point people to the highest quality, most authoritative information available, especially when it comes to an issue prone to misinformation. To determine what is borderline, we work with human evaluators and experts. We rely on humans for the same reason that human review is important for our content review prior to removal i.e. because a nuanced determination is often needed. The input of our evaluators and experts in turn helps to train the machine learning systems that generate recommendations to make them better at identifying borderline content and automatically reduce exposure.

Dangers of over-regulating ranking tools

YouTube has invested significantly in building algorithms and recommendations systems which raise authoritative content (as detailed above). These work effectively at dealing with major new misinformation challenges such as the war in Ukraine (see, for example, [recent research](#) from the Institute for Strategic Dialogue). We note that the Bill contains a number of provisions aimed at placing guardrails around platforms' use of algorithms and providing additional transparency or appeal rights on ranking decisions. While we support the aims and principles of these provisions at a high level, we strongly urge that any implementation avoids overly-prescriptive solutions which could force YouTube to implement a solution which limits its ability to offer UK users the protections, the technologies for which we have spent years developing.

User controls over YouTube tools

We give users control over their recommendations through YouTube settings in several ways:

- Users can view, pause, edit, or clear their watch history at any time through the YouTube history settings.
- Users can also clear their search history, remove individual search entries from search suggestions, or pause search history using the YouTube History settings.
- In-product controls enable users to remove recommended content - including videos and channels - from their Home pages and Watch Next.
- Users can disable autoplay in their setting or on any video watch page.
- Signed-in users can also choose to have their YouTube search and watch history deleted automatically after a certain period of time through their Google “My Account” settings.

We also ask users directly about their experience with individual videos and our recommendation systems using random surveys that appear on their homepage and elsewhere on the service. We use this direct feedback to fine-tune and improve these systems for all users.

Our approach to ranking content on Search

Even if we have not received a legal removal notice to remove an illegal URL from Search results, our ranking approaches provide protections against such content surfacing prominently. This “ranking-first” approach used by Google Search minimises the risk that users encounter illegal or legal but harmful content. As spammers and other bad actors are constantly evolving and adapting to evade counter-abuse technologies and techniques, we are continually improving our ranking approaches. This ongoing process helps to ensure that we are promoting authoritative and trusted content, while demoting content that could be harmful to users. The result is that such content is made far less prominent for queries without a clear intent to find it, in both web and image search.

In addition to the tools that we use to remove content, our Search systems are designed to prioritise what appears to be the most helpful content on a given topic, and not to surface content that violates our content policies. However, we also recognise that no system is perfect and we build this recognition into our approach. If, in spite of our

	<p>processes, policy-violating content is surfaced, we always aim to resolve it by improving our automated systems. This allows us to better deal with a particular issue that has been detected and to improve the approach for related queries and other searches. In some cases, we may also take manual action, which means that our team members review cases where policy-violating content surfaces and take manual action to block this content, in the limited and well-defined situations that warrant this.</p>
<p>Question 17: What other sanctions or disincentives do you employ against users who post illegal content?</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>YouTube: Account removals</p> <p>As well as the “three strikes” policy (explained in our response to Question 9 above), a YouTube user who has their channel or account terminated will be unable to use, own or create any other YouTube channels/accounts. Our Terms of Service explicitly prohibit users from using another channel to circumvent our enforcement measures.</p> <p>We ensure that sanctions are applied consistently and fairly through a combination of:</p> <ul style="list-style-type: none"> • the rigour with which we identify violative content (including a process for escalating borderline cases to our Trust & Safety team); • the application of the detailed approach set out in our Terms of Service; and • the inclusion within our sanctions processes of an appeal mechanism. <p>We also have a robust circumvention policy that relies on a variety of signals to help us determine if an account is likely to have been created by someone whose account was previously removed. This is also explained in our Terms of Service which make clear that when an account is “<i>turned off or restricted from using any YouTube features</i>”, the user is “<i>prohibited from using another channel to get around these restrictions</i>”.</p> <p>We do not have a repeat infringer policy for users whose content has been removed as a legal removal, given that more serious conduct is covered by the Community Guidelines process outlined above, and legal removals are specific, local law issues.</p> <p>Search: Sanctions</p> <p>In Search, website operators are not “users”. However, we still deploy ranking signals to monitor sites where we receive many valid legal removal requests. Sites that receive a sufficiently large volume of copyright notices, proportional to the size of the site, will receive a lower ranking on queries not explicitly seeking that site. We announced that signal in 2012 and, since then, we have rolled out similar signals for sites for which we</p>

receive valid delisting requests under defamation law, court orders, or our policy relating to counterfeit goods.

Other sanctions

There are also sanctions contained in Google's universal Terms of Service, which apply to signed-in users of other Google products (save where a service has a more specific process).

These terms state that:

"Google reserves the right to suspend or terminate your access to the services or delete your Google Account if any of these things happen:

- you materially or repeatedly breach these terms, service-specific additional terms or policies*
- we're required to do so to comply with a legal requirement or a court order*
- we reasonably believe that your conduct causes harm or liability to a user, third party, or Google — for example, by hacking, phishing, harassing, spamming, misleading others, or scraping content that doesn't belong to you.*

If you believe your Google Account has been suspended or terminated in error, you can appeal."

Question 18: Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry?

Is this response confidential? – Y/N (delete as appropriate)

Design features should be tailored to the service

Google believes there is no one-size-fits-all approach to designing services in a way that minimises the risk of harm. Different functionalities work well on different services, for different types of content. Different approaches and design features need to be used across our services, depending on the nature of the content on that service and its functionalities.

We explain above, in our response to [Question 12](#), that the use of automated technology such as hash matching is effective for detecting and removing CSAM, but it cannot identify illegal content where doing so requires a nuanced assessment of context; for example, hate speech, where it may be necessary to understand the intention of the person sharing the content to determine whether they have committed an offence.

We employ a safety by design approach, incorporating it into the core aspects of each of our services' functionalities, including in the design of algorithms, such as

recommendation algorithms. In our responses to [Question 19](#) and [Question 20](#) below, we describe specific functionalities and experiences which we believe are effective in preventing or reducing harm on our services.

We are keen to work with Ofcom and assist it to develop guidance on how the industry can incorporate a safety by design approach into the product development process.

Examples of how Google tailors safety features

To take an example, on services where content may be shared on sensitive issues, such as COVID-19 or elections, it is possible to design to prioritise information from established and trustworthy news sources including government health authorities, and established news media businesses. This is an approach we take on **YouTube**, which reduces exposure to misinformation and prevents harm. Evidence from independent studies shows that our efforts have been successful. For example, in relation to the war in Ukraine, the Institute for Strategic Dialogue [found](#) that "YouTube's results comprised almost exclusively media sources (38/40)".

Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact of harm from illegal content?

Is this response confidential? – Y / N (delete as appropriate)

As set out in our response to [Question 15](#), if we are aware of content that we have a legal obligation to remove, we disable access to it rather than simply mitigating the risk or impact of harm from it.

Google's approach

We design all of our services to mitigate the risk of harm. For example, and as explained above, our services are designed so that they prioritise what appears to be the most helpful content on a given topic, and not to surface content that violates our content policies. This design approach works alongside our review and removal processes.

As an example, in developing our policies, tools and features around self-harm or suicide queries or content, we consult with both internal and external experts in psychology, mental health, and related areas. These include not only academics and clinicians, but also practitioners who provide direct services to vulnerable populations. We know how important it is to increase awareness around help-seeking behaviours, while decreasing risk-taking and reducing stigma. This issue is complex and requires highly specialised expertise, which is why we have a dedicated Health team, led by Google's Chief Health Officer, with whom we work closely to inform our product design.

In addition, we have specific tools that we use to protect all users, including those that are most vulnerable.

On YouTube:

- We have specific functionalities to mitigate the risk of harm to children, as described in our answer to [Question 24](#) below.
- We also offer users an optional setting (Restricted Mode) that helps to screen out potentially mature content that some users may prefer not to view. This feature is designed to filter out graphic content that is permitted under Community Guidelines, but which is more appropriate for mature users.
- We anticipate problems before they emerge and adapt. As explained in our response to [Question 3](#) above, our Intelligence Desk monitors the news, social media and user reports from around the world to detect new trends, and works with the right teams to investigate and address them before they can become a larger issue.

On Search:

- Within Google's population of users, there are some for whom it is particularly critical that we get our results right: users in crisis. Some of our users turn to Google as their first or last resource after going through a traumatic event. For these reasons, we've done work to refine our systems to help improve the visibility of authoritative information, such as national hotlines and text services, in search results for queries that indicate a high intent of self-harm or suicide. When users in the UK express urgent intent around suicide, a feature will appear at the top of their Search results page. This feature surfaces phone numbers of the Samaritans that support users in "SOS" situations, free of charge. Our suicide hotline feature is Google's approach to connecting vulnerable users facing imminent harm with helpful and free resources immediately. We also ensure that support charities' websites appear at the top of the list of results. We also do not allow autocompletions on search terms related to suicide and self-harm.
- When users search for certain mental health conditions such as depression, we surface a knowledge panel with information from the NHS about clinical depression and give users the option to click through to more authoritative NHS information.
- We enforce [content policies on Search features](#), including Images and autocomplete, that aim to prevent the surfacing of dangerous content. We don't allow content on these features that could directly facilitate serious and immediate harm to people, such as content that promotes self-harm and eating disorders or content that provides instructions on committing suicide.

	<ul style="list-style-type: none"> • As with all of our search ranking systems, we're continually making improvements to ensure that we're providing people with the highest quality information possible, while also not showing people shocking or potentially harmful results that they do not explicitly seek. • In the UK, we provide SafeSearch as a default for users that we believe to be under 18 and make it available for all users, to help filter out explicit content. • As described above, if a user searches for CSAM, deterrence messaging is shown indicating the content is illegal.
<p>Question 20: How do you support the safety and wellbeing of your users as regards illegal content?</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>Google deploys a range of approaches, including tailored support to specific products and off-platform initiatives, to support the safety and wellbeing of users as regards illegal content.</p> <p>Google's approaches are not simply limited to removing or preventing access to illegal content, but we aim to support and empower users to navigate the internet safely and improve their wellbeing.</p> <p>Search</p> <p>As explained above in our response to Question 12, we use deterrence messaging when users search for illegal CSAM content. We surface warnings for those that are searching for CSAM, which include links to the "Stop It Now" campaign for users that are concerned about their own feelings. We also provide links to organisations that can support victims of child sexual abuse and exploitation; for example, Childline in the UK.</p> <p>As explained in our response to Question 19, we also offer tools that directs users at risk of suicide to resources that could offer immediate help, such as the Samaritans. We have specific policies to prevent harmful content from appearing in search features such as autocomplete or snippets for this type of search.</p> <p>Independent academics have studied our results for suicide-related queries, and published research that found that Google performs better than other search engines when it comes to handling suicide-seeking queries. They found that other major search engines returned more harmful URLs when compared to Google, and they were also less likely to display help messaging at the top of the search results page.</p> <p>We have recently updated Search to use our AI Multitask Unified Model (MUM) to automatically and more accurately detect personal crisis searches in 75 languages in</p>

order to show the most relevant information when our users need it. MUM is a powerful AI model that is capable of deeper and more nuanced language understanding, making it more adept at identifying when a query is about a crisis. When we detect a query is about a crisis, we are able to improve the ranking of trustworthy information and more reliably surface actionable information such as local hotlines.

Every improvement to Google Search undergoes a rigorous evaluation process to ensure we're providing more relevant, helpful results. These improvements have only been made possible by advanced AI and we continue to invest in this area.

YouTube

Under our [suicide and self-injury policy](#), we prohibit content that promotes self-harm or is intended to shock or disgust viewers. We remove content promoting or glorifying suicide, content providing instructions on how to self-harm or die by suicide, and content containing graphic images of self-harm posted to shock or disgust viewers. For users looking for videos to aid suicide in the UK, we also offer messaging directing them to the Samaritans.

While pornography is not illegal, accessing it is illegal or heavily limited for children under the age of 18. We have a number of protections in place to prevent children from accessing this content. For example, we gate for under 18s videos containing nudity and sexually suggestive content, content which shows adults participating in dangerous activities that minors could easily imitate and cause injury, and violent and graphic or vulgar language. In our response to [Question 24](#), we describe the methods that we use to establish if a user is over the age of 18.

We have introduced a number of features to promote digital wellbeing and are working hard to ensure we understand how we can best protect our users by carrying out research and updating our products. We are engaged in numerous initiatives to support the wellbeing of our users, both directly through our services and off-platform. Digital Wellbeing is an area of tech use that is very personal, and there is not a one-size-fits-all approach. We have created a set of [Digital Wellbeing tools](#) that are designed to help users find the right balance with technology for themselves.

Off-platform

Recent studies⁶ have highlighted that digital media use can help teens communicate with peers and family, seek helpful resources if they are experiencing distress, and find opportunities for learning and entertainment that can help combat isolation. This is why we have established long-standing partnerships in the UK aimed at strengthening media literacy:

- We partner with Parent Zone on "[Be Internet Legends](#)", the only PSHE-accredited online safety programme for 7-11 year olds in the UK, which has reached over 71% of primary schools in the UK since launching in March 2018.
- In partnership with the Institute for Strategic Dialogue, Google and YouTube have launched a programme called "[Be Internet Citizens](#)" (BIC) to equip 13-15 year olds with the media literacy skills to experience the internet in a safe and positive way. Since 2017, BIC has reached an estimated 80,000 young people across the UK, while over 850 teachers and youth workers have been trained to deliver the curriculum independently. Based on feedback, 84% of young people having gone through the programme feel confident they would know what to do if they encountered hate speech online. This year, we have launched two series of [YouTube Reframe](#) to scale the BIC curriculum and use our creators to present these topics to all 13-15 year olds who use YouTube. Reframe has already reached more than 300,000 views.

We also support a wide number of organisations that work to prevent harm and support victims and survivors. Google.org has awarded a grant to support the development of a hub of excellence in suicide prevention and the online environment, incorporating research and the development of evidence-based resources and guidelines for both professionals and the public. This is part of a wider programme of support to NGOs working to prevent child sexual abuse and support children to improve their digital skills.

Question 21: How do you mitigate any risks posed by the design of algorithms that support the function of your service (e.g. search engines, or social and

Is this response confidential? – Y / N (delete as appropriate)

Across our products and services

Algorithms are an integral part of how our services function and meet the needs of our users. We ensure that they are designed to prioritise access to the most helpful information on our services.

⁶ For example, [this paper](#) published by UNICEF (Stoilova, M., Livingstone, S., and Khazbak, R. (2021) Investigating Risks and Opportunities for Children in a Digital World: A rapid review of the evidence on children's internet use and outcomes. Innocenti Discussion Paper 2020-03. UNICEF Office of Research – Innocenti, Florence) and [this paper](#) published in the Frontiers in Digital Health journal (Pretorius C, Coyle D. Young People's Use of Digital Tools to Support Their Mental Health During Covid-19 Restrictions. Front Digit Health. 2021 Dec 1;3:763876. doi: 10.3389/fdgth.2021.763876. PMID: 34927133; PMCID: PMC8671300.

content recommender systems), with reference to illegal content specifically?

Google has developed cross-product [artificial intelligence principles](#), which set out our commitment to developing technology responsibly.

However, algorithms are not infallible. A [recent report](#) by the Government's independent advisory body, the Centre for Data Ethics and Innovation (CDEI), noted that algorithms are "generally poor at contextual interpretation", making their deployment to identify most forms of illegal content challenging. For example, algorithms would struggle to distinguish between content from a terrorist organisation glorifying violence and content from a journalistic or human rights organisation documenting such violence.

As reflected in our AI principles (linked above), we seek to mitigate this sort of risk by, among other things, reviewing our machine learning approaches to reduce the risk of unintended algorithmic bias in our trust and safety systems.

YouTube

Recommendations on YouTube are designed to minimise the chances that users will see problematic content. Our Community Guidelines set the rules of the road on YouTube, and a combination of people and machines help us remove more violative content than ever before. That said, there will always be content on YouTube that brushes up against our policies, but doesn't quite cross the line. So we work to raise authoritative voices on YouTube and reduce the spread of borderline content and harmful misinformation. We are seeing great progress. Authoritative news is thriving on our site. In 2019, we launched over 30 different changes to reduce recommendations of borderline content and harmful misinformation. The result was a 70% average drop in watch time of this content coming from non-subscribed recommendations in the U.S. by the end of 2019.

We use classifiers to identify whether a video is "authoritative" or "borderline". These classifications rely on human evaluators who assess the quality of information in each channel or video (as explained in our response to [Question 16](#)). We rely on certified experts, such as medical doctors when content involves health information.

Search

We use automated systems to deliver the most relevant and reliable information. These systems consider many factors, including the words in a query, the content of pages, the expertise of sources, and the user's language. We have a [rigorous testing process](#) to ensure that our automated systems return high quality results.

Moreover, we put all possible changes to Search through a rigorous evaluation process (as described above in response to [Question 4](#)) to analyse metrics and decide whether to implement a proposed change. Data from these evaluations and experiments go through a thorough review by experienced engineers and search analysts, as well as other legal and privacy experts, who then determine if the change is approved to launch.

We gather data in multiple ways: we have human raters who look at side-by-side comparisons and tell us which results they prefer, we do A/B testing; and we survey users to ask about new features. Every change to Search goes through a launch process before it's approved. Of the proposed changes this past year, many never went live, because unless we can show a change is actually better for users, we don't launch it. In 2021, we ran over 750,000 quality tests that resulted in more than 5,000 improvements to Search.

A crucial part of ensuring that our automated systems work effectively is our Search Quality Rater programme, which is described in more detail in our response to [Question 14](#) above.

Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?

Is this response confidential? – Y / N (delete as appropriate)

There are a range of age assurance tools available to platforms, including those that we use (which are described in our response to [Question 24](#) below). These tools have been designed in accordance with the principles and requirements of the Age Appropriate Design Code promulgated by the ICO.

We are aware that there are a number of third party tools available in the market that can help with the age verification of users. We monitor the development of these tools and are always looking for opportunities for collaboration. We have not undertaken an assessment of the individual accuracy of each of these tools or their costs. At this stage, our focus is on developing our own solutions, which (as described in our response to [Question 24](#)) based on our age inference model, can help us to obtain a helpful and proportionate understanding of the age of our users at scale while minimising the use of data and the security risks that may result with the use of third party products. This market, and considerations about the effectiveness, costs, privacy minimisation and security of the technology available, will likely evolve in the next few years.

We are funding research to help us and the wider sector understand what are the different technologies available and how users, both children and adults, perceive and engage with these tools. In particular, we are funding research by the Family Online Safety Institute (FOSI) to help us understand how users interact with different technologies and what are the tradeoffs that they face in three markets - UK, France and US. We expect the results from this research later in 2022.

<p>Question 23: Can you identify factors which might indicate that a service is likely to attract child users?</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>We build services specifically designed for young users. These include YouTube Kids, Supervised Experiences on YouTube, Google Kids Space and Expert Approved apps on Play. We also offer Family Link to enable parents and carers to open and supervise accounts for children under the age of 13 (as explained in our response to Question 24 below).</p> <p>More generally, Google builds services that are useful to our users: they help them to learn, explore, communicate and have fun. These services include Search, YouTube, Play, Maps, Google Workspace, Google for Education, Android, Drive and Photos. We assume, and expect, that young people will benefit from these products.</p> <p>We assume that children will not be attracted to Google’s business directed services.</p>
<p>Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>Google is committed to ensuring that children and teens have appropriate experiences when using our products and services, and understanding the age of our users forms a part of this.</p> <p>As we discuss below, we use various tools to verify the age of users or for age assurance purposes. We also use other tools and services - some of which are product-specific - to limit access to content that is inappropriate for children.</p> <p>Our approach to age assurance</p> <p>In the UK, we utilise age inference technology to provide more age appropriate experiences for our users. Specifically, we deploy machine learning models that help to provide an additional level of assurance with regard to the veracity of a user’s declared age, or to provide an indication as to whether or not a user is an adult where the user has not declared their age. This in turn helps us to determine the appropriate levels of protection to apply for our users. These models use a variety of signals, such as the types of sites a user is searching for or the categories of videos that they have watched on YouTube, as well as indicators like the longevity of an account, to make a determination on age. For example, searches for mortgage lending sites or tax assistance would be signals that are indicative of the likelihood that a user is an adult. We are constantly iterating the model to improve its accuracy. We do not collect any additional data for this purpose, but rather utilise only the data that is available in accordance with the user’s privacy settings.</p> <p>We require users to provide additional verification of their age in limited circumstances, for example, we might require age verification if a user is trying to access age-restricted</p>

content or services, and we cannot otherwise establish with sufficient certainty that they are an adult, or if our age assurance process has classified the user as under 18 and they wish to access age-restricted content.

We use the following tools to obtain additional verification of age:

- **Government ID:** A user can submit an electronic copy of a valid, government-issued ID that shows the user's date of birth. That submission is then reviewed and approved (normally within 24 hours). To protect users, we then delete the copy after we have validated the user's date of birth. The user can cover the national identification number on their ID to keep it confidential.
- **Credit card verification:** A user can also verify their age by providing a valid credit card. The user is not billed as part of this process but may see a small authorisation on their account from the request, which is subsequently removed.

Limiting access to inappropriate content

We use the following tools and services to limit access to content that is inappropriate to children:

- **Restricting access to under 18s:** If a user provides us with their date of birth, and this indicates that they are under 18, or if our age assurance model indicates that they are under 18, then they will not have access to age-restricted content.
- **Appropriate advertising:** We disable ads personalisation for children under 18. We have also expanded safeguards to prevent age-sensitive ad categories from being shown to teens. Our goal is to ensure we're providing additional protections and delivering age-appropriate experiences for ads on Google.
- **Family Link:** All users under the age of 13 have to have an account managed with Family Link. We also offer Family Link parental controls by default on all Android devices for parents that want to set up content restrictions for their child. This functionality helps parents stay in the loop and guide their children as they explore and enjoy the internet. It allows parents to set and tailor digital ground rules that work for their family (including the ability to set screen limits, manage their apps, and lock their device). It also includes tips for families to help parents guide their children to make smart choices when using their own devices.

- **Google Kids Space:** This is an Android tablet experience which allows access to appropriate content. Children can access apps, books, and videos that are targeted to their age and interests. Parents can manage the experience; for example, by managing the content that can be seen and setting screen time limits.
- **On Search:**
 - **SafeSearch:** By default, on Google Search, we turn on SafeSearch for users under the age of 18. Our SafeSearch feature helps filter explicit results from Google Search results, even when they might be relevant for the query. While these algorithms will never be completely accurate, turning on SafeSearch helps to filter explicit content, like pornography, from Google search results. More generally, we also more prominently surface digital wellbeing features, and provide safeguards and education about commercial content.
 - **Removal:** While we already provide a range of removal options for people using Google Search, children are at particular risk when it comes to controlling their imagery on the internet. We therefore have a policy in place that enables anyone under the age of 18, or their parent or guardian, to request the removal of their images from Google Image results.
- **On YouTube:**
 - We age-restrict content on YouTube that does not violate our Community Guidelines but that may still not be appropriate for viewers under 18. For example, videos that contain vulgar language, violent or gory content, depictions of adults engaged in dangerous behaviour that might be emulated by minors, and adults consuming products that are not legally available to children (for example, alcohol or legal drugs). We continue to build on our approach of using machine learning to detect content for review, by developing and adapting our technology to help us automatically apply age-restrictions. Viewers attempting to access age-restricted videos on most third-party websites will be redirected to YouTube where they must sign in as an 18+ user to view it.
 - **YouTube Kids:** YouTube Kids offers a set of parental controls to customise their child's experience. Parents can decide what content to make available for their child to watch, set a timer to control screen time, block videos or channels, and more.

- **YouTube supervised experiences:** We offer parents the ability to create supervised accounts using Family Link. Supervised experiences allow children under the age of consent to access YouTube with parents choosing the right content setting for their children: “Explore”, “Explore More”, or “Most of YouTube”. The YouTube supervised experience looks much like YouTube’s flagship platform but with additional safety features.
- **Age restrictions on Play:** On Play, we prevent users who we know to be under the age of 18 from browsing or downloading apps classified as 18+.

A future regulatory framework on age assurance

We believe that age assurance measures should complement parental tools that help put parents at the centre of deciding what is best for their children and families. These measures should build on robust product design and clear policies to ensure that users, and children in particular, have a safer and more enriching online experience.

No age assurance mechanism is completely accurate, and the more accurate the mechanism, the more intrusive it can be. Ensuring that we implement age-appropriate safeguards, while at the same time ensuring that our services respect privacy and remain accessible remains a complex challenge. We are committed to tackling this challenge, but this will require a coordinated industry approach.

We are concerned that potential requirements under the Bill, which demand separate consideration to how individual pieces of content might affect children in different age groups, could be overly complex and unworkable. Mandating age-specific experiences for different age groups not only risks adding excessive complexity to services (as well as being potentially impossible to achieve technically), but it could also fail to take into account the differences in development and maturity that can occur during teenage years.

We think a better approach to tailoring technology to ensure it is relevant to specific age groups is to make flexible parental controls available on services, such as those described above. These tools put parents in control of the content and experience that their children can access, giving them the flexibility to choose what is right for their children and their families, taking into account different maturity levels and developmental abilities.

We also share the concerns of many that age restrictions may lead to the denial of some digital services to children, depending on the age limit. This could present a barrier to the educational development of teenagers, and prevent access to support systems, as well as limiting their ability to express their views and exercise their right to free speech.

<p>Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>We use a combination of policies, and tools that help us to enforce them, to prevent children from accessing some of the content available on our services.</p> <p>Our policies are clear from our universal Terms of Service, which include clear and accessible information about our age requirements:</p> <p><i>If you're under the age required to manage your own Google Account, you must have your parent or legal guardian's permission to use a Google Account. Please have your parent or legal guardian read these terms with you.</i></p> <p><i>If you're a parent or legal guardian, and you allow your child to use the services, then these terms apply to you and you're responsible for your child's activity on the services.</i></p> <p>These policies are backed-up by various tools, as described above in our response to Question 24, which we apply as appropriate to children aged under 13 and children aged under 18. For example, if a user is attempting to access age-restricted content on YouTube or 18+ apps, and our systems are unable to confirm that a user is likely to be above the age of 18, they will be unable to watch or download the age restricted content.</p>
<p>Question 26: What information do you have about the age of your users?</p>	<p><i>Is this response confidential? – Y / N (delete as appropriate)</i></p> <p>Despite our extensive and robust tools, we agree with the view that the requirements to use identity verification and age assurance measures could lead to excessive collection of users' data, including children's data. We would therefore welcome further guidance on how services are expected to uphold their responsibilities to both user safety and privacy.</p> <p>The range of age assurance, age inference and (where necessary) age verification tools we use to ascertain the age of our users are described in our response to Question 24 above.</p> <p>Currently, Google will retain some data as necessary to meet legal and regulatory requirements. We do not share information ascertained from these tools with third parties. Where a user uploads a government ID, we delete the copy of the document after we have validated the user's date of birth.</p>

Question 27: For purposes of transparency, what type of information is useful/not useful? Why?

Is this response confidential? – Y / N (delete as appropriate)

Google's approach to transparency

We believe it is important to be clear and transparent with users, researchers and the wider public about the measures we are taking to remove harmful and illegal content from our services and their effectiveness. We have considered carefully where we can inform individuals about content removals and we believe that Google is an industry leader in transparency.

We focus on being clear to users about:

- requests by government to remove illegal content;
- how we have enforced our service policies; and
- how we use algorithms to rank content and recommendations.

However, our approach to transparency varies across services and there can be no one-size-fits-all approach. For example, our introduction of VVRs on YouTube (discussed in our response to [Question 2](#)), shows how many times content has been viewed before it is removed for breaching our policies. As noted in that response, we see these VVRs as our “North Star” for measuring our progress in combating harmful content and we believe that sharing these rates with the public is an important way to create accountability. However, this approach is specifically designed for the YouTube platform and would not necessarily be an appropriate solution for other platforms.

YouTube

We have provided detailed information about the types of data we publish on removal of illegal content and how we remove this content in our responses to [Question 2](#) and [Question 8](#).

On the [How YouTube Works](#) site, we explain our approach to algorithms and the user controls we offer:

“We're constantly testing, learning and adjusting to recommend videos that are relevant to you. We take into account many signals, including your watch and search history (if enabled) as well as the channels that you've subscribed to. We also consider your context, such as your country and time of day. For example, this helps us show you locally relevant news. Another factor that YouTube's recommendation systems consider is whether others who clicked on the same video watched it to completion – a sign that the

video is higher quality or enjoyable – or just clicked on it and shortly after starting to view the video, clicked away.”

Search

Since Google launched its first Transparency Report in 2010, we have been committed to extensive transparency, which is one of our core values:

- For users and webmasters, we make information available to explain how Search algorithms sort through hundreds of billions of webpages to present the most relevant, useful results in a fraction of a second. These include:
 - A [How Search Works website](#) that explains concepts like crawling, indexing, and ranking – as well as our testing and evaluation processes and spam protections.
 - Our [Search Quality Rater Guidelines](#), which explain how we use human raters to help make sure Search is returning relevant results from the most reliable sources available.
 - Annual webspam reports ([2020](#), [2019](#), [2018](#), [2017](#), [2016](#)) which provide an overview of how our systems detect 40 billion spammy pages a day.
 - In-product features like [About this Result](#) which we’ve expanded to show searchers information about some of the [most important factors](#) used by Google Search to connect results to their queries.
- We have a long history of supporting transparency reporting. Currently, we issue reports on violations related to:
 - [CSAM](#), including the number of CyberTipline reports, the volume of content reported to NCMEC, and the number of URLs de-indexed.
 - [Copyright](#), including the number of URLs requested to be delisted, the number of specified domains, the number of copyright owners (individuals or entities), and the number of reporting organisations acting on behalf of copyright owners.
 - [Counterfeit goods](#), including the number of URLs requested to be delisted, the number of specified domains, the number of brand owners (individuals or entities), and the number of reporting organisations acting on behalf of brand owners.
- We also report extensively on [government requests](#) to remove content. This report is broken down by the product, the country of the request, and the reason for the removal request.
- In addition, Google Search provides bespoke transparency reporting in response to individual local laws, including:
 - [European privacy law](#), otherwise known as the “right to be forgotten”.
 - [South Korea’s Network Act and the Telecommunications Business Act](#), focused on illegal sexual content such as CSAM and NCEI.

In addition to these transparency reports, Google Search also provides additional transparency via the [Lumen database](#). When we receive a legal removal request for Search, we transmit a copy to Lumen, a project run out of Harvard University. Lumen then partially redacts and publishes the request on its website. When a Google search results page is affected by a legal removal, we link to Lumen’s copy of the request:

In response to a legal request submitted to Google, we have removed 2 result(s) from this page. If you wish, you may [read more about the request](#) at [LumenDatabase.org](#).

Transparency reporting for Search is necessarily different from transparency reporting for other products, such as user-generated content platforms. For example:

- Google Search does not report in some areas that user-generated content platforms often report on, such as account-level enforcement. Search does not enforce against individual users, using methods such as a graduated system of strikes, account bans, feature blocks (for example). The reason for this is that search engines generally lack a direct, contractual relationship with content creators. Often, the only information Google possesses about a given website is what is available to anyone with a web browser. Therefore, obligations requiring reporting metrics on account-level enforcement make little sense for a product like Google Search.
- Conversely, Google Search needs to invest in reporting in certain areas that are not necessarily priority issues for other types of service. For example, Google Search has processed [billions of DMCA requests](#) and it has invested in extensive transparency reporting for content delistings due to copyright. Such reporting is appropriate to search engines, given the volume of copyright-related requests that they receive and process. However, such extensive transparency reporting may not be as necessary or even relevant for other types of products and services, for which copyright may be a relative non-issue. Some hosted platforms process only a dozen or so requests a year, given the nature of their service, or do not report on DMCA at all.

In short, search engines need bespoke transparency reporting obligations that reflect the nature of these products and the way that they are used.

We would add the following additional points to consider when developing requirements around transparency for a product like Search:

- **Value of regulatory alignment**
 - We recognise the value of the DRCF in setting out the potential actions regulators should consider, for example, to develop and shape solutions to algorithmic auditing. As noted in their [recent paper](#) on this topic,

increased coordination amongst regulators may facilitate the trustworthy use of algorithms and provide the regulatory clarity needed to stimulate innovation by ensuring proportionate compliance costs and removing unnecessary burdens.

- **Transparency is an adversarial space**

- Exposing an algorithm's code is not a practical means of boosting accountability and brings with it a real risk of making our systems harder to protect.
- We appreciated the need for these safeguards the hard way. Back in 1999, Google's founders published a seminal [paper](#) on PageRank, a key innovation in Google's algorithm, which discussed the parameters used to determine Search rankings. Once that paper was published, spammers used that information to try to manipulate results. Even today, our systems discover 40 billion spammy pages every single day.

- **Search is both extremely complex and incredibly dynamic**

- Search is a complex knowledge and information product. It is not a single algorithm and it is far from static. After URLs are crawled and indexed, our ranking systems are sorting through hundreds of billions of webpages and other content in our Search index to present the most relevant, useful results in a fraction of a second. No single algorithm could accomplish this task.
- Search is also far from static. To keep Google Search running and meeting the needs of users, we undergo rigorous testing and evaluation. In 2021 alone, we ran more than 700,000 quality tests, 11,500 live experiments, and 72,000 side-by-side experiments, resulting in more than 5,000 improvements to Search.
- Exposing the code of a single algorithm does not reveal much. It takes significant technical expertise, resourcing, and time for in-depth analysis. For example, the [FTC needed four years](#) to analyse a single algorithm.

Areas where transparency could negatively affect users

Exposing an algorithm's code can make our systems harder to protect by creating opportunities for bad actors to exploit them, such as through hacking and fraud. We have a responsibility to protect our consumers and our systems from these security risks.

We recognise that regulators like Ofcom may have a legitimate desire to access more information about how our content moderation systems and other technologies work in

practice. However, to the extent that this information could then be made public, either by Ofcom directly or through requirements it imposes to publish transparency reports in full, it is important to acknowledge the risk that:

- the information could be used by bad actors to game systems and evade our content moderation efforts;
- commercially sensitive information and trade secrets could be exposed; and
- sensitive user information could be disclosed, adversely affecting user privacy.

Information that could be shared in transparency reports required by Ofcom

As explained above, Google already publishes transparency reports and other material, including on our content moderation, as part of our long-standing commitment to providing an open, transparent relationship with those who use our services.

However, providing more sensitive information about our technology - such as detailed insight into algorithms - must be balanced against the potential risks to the security and integrity of our products. Any requirements to do so should be accompanied by proportionality safeguards such as (a) only exercising these powers after use of regular RFI / interview powers have proved insufficient; (b) a convincing statement of reasons as to why it is necessary to exercise these powers in the case at hand; (c) a duty to limit requests to the minimum necessary; and (d) a duty to take account of business disruption caused by such a request.

Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?

Is this response confidential? – Y / N (delete as appropriate)

In our response to Question 3 above, we set out the tailored approach to tools and policies we currently deploy on our services to ensure user safety. Measures and technology to mitigate risk and harm often evolve. We welcome the opportunity to brief Ofcom in the future as we introduce new tools and measures to mitigate risk and harm from illegal content.

Please complete this form in full and return to OS-CFE@ofcom.org.uk