# UK Online Safety – OFCOM Call for Evidence

# RESPONSE ON BEHALF OF META PLATFORMS IRELAND LTD

We would like to thank OFCOM for the opportunity to provide information in response to this call for evidence.

Over the last five years, Meta Platforms Ireland Ltd has supported the UK Government's development of the Online Safety framework through evidence sessions, written submissions, ministerial discussions, and multi-stakeholder in-person technical sessions. We share the UK Government's stated policy objectives, to make the internet safer while protecting the vast social and economic benefits it brings to billions of people each day. At Meta, we have eighteen years' experience in tackling online safety issues through establishing policies, building tools and technologies, and producing guides and resources, all in partnership with experts both within and outside our company.

We are pleased to use that experience to contribute to OFCOM's preparation for its work as regulator. As set out in our responses below, we are happy to engage further with OFCOM to identify additional information that may help to inform OFCOM's thinking. We look forward to continuing our constructive relationship.

## PRELIMINARY QUESTION

### 1. Please provide a description introducing your organisation, service or interest in Online Safety.

Meta Platforms Ireland Ltd's ("Meta") mission is to give people the power to build community and bring the world closer together.

We build technology that helps people connect, find communities, and grow businesses. Our useful and engaging products enable people to connect and share with friends and family through mobile devices, personal computers, virtual reality (VR) headsets, wearables, and in-home devices. We also help people discover and learn about what is going on in the world around them, enable people to share their opinions, ideas, photos and videos, and other activities with audiences ranging from their closest family members and friends to the public at large, and stay connected everywhere by accessing our products.

Meta provides a number of different services and functionalities to users, as set out below. For users in the UK, these services are provided by Meta Platforms Ireland Ltd.

Facebook: Facebook helps give people the power to build community and bring the world closer together. It's a place for people to share life's moments and discuss what's happening, nurture and build relationships, discover and connect to interests, and create economic opportunity. They can do this through the features within Facebook: Feed, Stories, Groups, Watch, Marketplace, Reels, Dating, and more.

Instagram: Instagram brings people closer to the people and things they love. Instagram Feed, Stories, Reels, Video, Live, Shops, and messaging are places where people and creators can express themselves and push culture forward through photos, video, and private messaging, and connect with and shop from their favourite businesses.

Facebook and Instagram services are available to any person in the UK aged 13 years and older, subject to limited exceptions set out in the Terms of Service (such as where users are prohibited by law from using these services). The types of content that can be published and consumed using the Facebook and Instagram services are described in the Terms of Service and the Community Standards (Facebook) (https://transparency.fb.com/en-gb/policies/community-standards/) and Community Guidelines (Instagram) (https://help.instagram.com/477434105621119). The Facebook and Instagram services are social media services.

Messenger: Messenger is a simple yet powerful messaging application for people to connect with friends, family, groups, and businesses across platforms and devices through chat, audio and video calls, and Rooms.

## RISK ASSESSMENT AND MANAGEMENT

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

Meta aims to create safe and trusted platforms, where people can feel free to express themselves. Our terms and policies do not allow people to post content that is against the law or encourages criminal behaviour. We also do not allow for bullying or harassment in any form. In order to achieve the balance between freedom of expression and protection from harmful and illegal content, we take a multi-faceted approach to addressing potentially criminal or harmful activity on our platforms.

First, we maintain globally applicable standards – Facebook's Community Standards and Instagram's Community Guidelines – that define what is and isn't allowed on our services. These standards apply uniformly to content worldwide and are integral to protecting expression and enhancing personal safety on our services.

Many of our standards focus on content that is or is likely to be illegal in the UK, including content that would be classified as 'priority illegal content' under the Bill – for example, see the 'Violence and Incitement' section of Facebook's Community Standards (available here: https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/). Our standards address the types of potentially harmful content that are of greatest concern or are seen most commonly on our platforms. In addition, our standards prohibit a wide range of objectionable or harmful content that is not necessarily illegal in the UK, including content that is considered hate speech, graphic violence, spam, misinformation, bullying or harassment.

Our standards are created by global teams with a wide array of backgrounds and expertise, including those who have dedicated their careers to issues like child safety, hate speech, and terrorism. We regularly seek input from outside experts and organisations to help balance the different perspectives that exist on free expression and safety, and to better understand the potential impacts of our policies on different communities globally. Our reviewers enforce these standards using comprehensive guidelines, in an effort to ensure that decisions are as consistent as possible.

Second, across our various platforms we action[1] millions of pieces of content per day, through both human review and automation (discussed further below). To track our progress and demonstrate our continued commitment to making Facebook and Instagram safe and inclusive we publish the Community Standards Enforcement Report (CSER) on a quarterly basis. The latest report (available here:

[1] As explained in Question 10, taking action on content could include removing a piece of content from Facebook or Instagram, covering photos or videos that may be disturbing to some audiences with a warning, or disabling accounts.

https://transparency.fb.com/data/community-standards-enforcement/) is from August 2022 and shares updated metrics for the reporting period from April to June 2022, detailing the prevalence of user accounts found to be in violation of our policies and our progress in preventing and/or taking action against content that violates our policies.

As set out above, content that violates our policies may include both legal and illegal content, and we do not conduct a legal review for every item of content that we action. As such, the information provided below (taken from the August 2022 report) is not limited to illegal content. It covers all user generated content (text, video, image, user interaction, excluding comments) that violates our policies. Please note that these are global figures.

Facebook:

Prevalence[2] of Facebook content that violates our policies, focusing on categories that are likely to overlap with illegal content:

- Adult Nudity and Sexual Activity: 0.04%
- Bullying and Harassment: 0.08-0.09%
- Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation of Children: We cannot estimate prevalence for child endangerment right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
- Dangerous Organisations: Terrorism and Organised Hate: In Q2 2022, the upper limit was 0.05% for violations of our policy for terrorism on Facebook. We cannot estimate prevalence for organised hate right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
- Hate Speech: 0.02%
- Regulated Goods: Drugs and Firearms: 0.05%
- Suicide and Self-Injury: 0.05%
- Violence and Incitement: 0.03%
- Violent and Graphic Content: 0.04%

Items of Facebook content actioned in the last four quarters for each of the above categories:

| **Category** | **Content Actioned Q2 2022** | **Content Actioned Q1 2022** | **Content Actioned Q4 2021** | **Content Actioned Q3 2021** |
|---|---|---|---|---|
| Adult Nudity and Sexual Activity | 38.4m | 31m | 27.3m | 34.7m |
| Bullying and Harassment | 8.2m | 9.5m | 8.2m | 9.2m |

[2] Prevalence is the estimated number of views that showed violating content, divided by the estimated number of total content views on the relevant service, sampled globally. Whilst we consider prevalence to be a good indicator when considered globally, this is a difficult metric to apply at national level. Put simply the error margin increases when the sampling area decreases.

| | | | | |
|---|---|---|---|---|
| Child Endangerment: Nudity and Physical Abuse Sexual Exploitation | Child Sexual Exploitation: 20.4m<br>Child Nudity and Physical Abuse: 1.9m | Child Sexual Exploitation: 16.5m<br>Child Nudity and Physical Abuse: 2.1m | Child Sexual Exploitation: 19.8m<br>Child Nudity and Physical Abuse: 1.8m | Child Sexual Exploitation: 21.2m<br>Child Nudity and Physical Abuse: 1.8m |
| Dangerous Organisations: Terrorism and Organised Hate | Terrorism: 13.6m<br>Organised Hate: 2.3m | Terrorism: 16.1m<br>Organised Hate: 2.5m | Terrorism: 7.7m<br>Organised Hate: 1.6m | Terrorism: 10.6m<br>Organised Hate: 2m |
| Hate Speech | 13.5m | 15.1m | 17.4m | 22.3m |
| Regulated Goods: Drugs and Firearms | Drugs: 3.9m<br>Firearms: 1.6m | Drugs: 3.3m<br>Firearms: 1.2m | Drugs: 4m<br>Firearms: 1.5m | Drugs: 2.7m<br>Firearms: 1.1m |
| Suicide and Self-Injury | 11.3m | 6.8m | 6.1m | 8.5m |
| Violence and Incitement | 19.3m | 21.7m | 12.4m | 13.6m |
| Violent and Graphic Content | 45.9m | 26.1m | 25.2m | 26.6m |

Instagram:

Prevalence of Instagram content that violates our policies, focusing on categories that are likely to overlap with illegal content:

- Adult Nudity and Sexual Activity: 0.02% to 0.03%
- Bullying and Harassment: 0.04-0.05%
- Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation: We cannot estimate prevalence for child endangerment right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
- Terrorism and Organised Hate: In Q2 2022, the upper limit was 0.05% for violations of our policy for terrorism on Instagram. We cannot estimate prevalence for organised hate right now. We will continue to expand prevalence measurement to more areas as we confirm accuracy and meaningful data.
- Hate Speech: 0.01-0.02%
- Regulated Goods: Drugs and Firearms: 0.05%
- Suicide and Self-Injury: 0.05%
- Violence and Incitement: 0.01-0.02%
- Violent and Graphic Content: 0.01-0.02%

Items of Instagram content actioned in the last four quarters for each of the above categories:

| **Category** | **Content Actioned Q2 2022** | **Content Actioned Q1 2022** | **Content Actioned Q4 2021** | **Content Actioned Q3 2021** |
|---|---|---|---|---|
| Adult Nudity and Sexual Activity | 10.3m | 10.4m | 11.3m | 10.9m |
| Bullying and Harassment | 6.1m | 7m | 6.6m | 7.8m |
| Child Endangerment: Nudity and Physical Abuse Sexual Exploitation | Child Sexual Exploitation: 1.2m<br>Child Nudity and Physical Abuse: 480k | Child Sexual Exploitation: 1.5m<br>Child Nudity and Physical Abuse: 601k | Child Sexual Exploitation: 2.6m<br>Child Nudity and Physical Abuse: 983k | Child Sexual Exploitation: 1.6m<br>Child Nudity and Physical Abuse: 527k |
| Dangerous Organisations: Terrorism and Organised Hate | Terrorism: 1.9m<br>Organised Hate: 449k | Terrorism: 1.5m<br>Organised Hate: 481k | Terrorism: 906k<br>Organised Hate: 332k | Terrorism: 685k<br>Organised Hate: 306k |
| Hate Speech | 3.8m | 3.4m | 3.8m | 6m |
| Regulated Goods: Drugs and Firearms | Drugs: 1.9m<br>Firearms: 215k | Drugs: 1.8m<br>Firearms: 151k | Drugs: 1.2m<br>Firearms: 195k | Drugs: 1.8m<br>Firearms: 154k |
| Suicide and Self-Injury | 6.4m | 5.1m | 7.8m | 3.5m |
| Violence and Incitement | 3.7m | 2.7m | 2.6m | 3.3m |
| Violent and Graphic Content | 10.2m | 6.1m | 5.5m | 10.7m |

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

We take a number of steps to assess and mitigate the risk of harm to our users from content that violates our standards (which, as noted above, overlaps with various types of illegal content) through the entire product development cycle.

In relation to product development, we build our products and continually update them with safety and integrity in mind. We embed teams focusing specifically on safety and integrity issues directly into our wider product development teams, allowing us to address issues at the development stage. We also offer integrity tools, built centrally, to individual product teams to allow them to build in preventative safeguards at the outset of the product development process.

Before launch, new products generally go through an Integrity Review, a cross-functional (XFN) process which evaluates product changes on integrity criteria prior to launch. In this process, products are reviewed against a set of integrity standards to help us provide a positive experience for users. As part of this process, we systematically and repeatedly bring together experts from across the company, including data scientists, safety experts and engineers. The process helps us identify and anticipate potential abuses and build in mitigations by design, prior to a product being launched.

After launch, we continue to monitor the potential impact of our products, including by looking at integrity metrics to ensure products are best serving our community. Teams are expected to establish ways to measure and track negative experience and enforcement effectiveness. We also have an XFN team that analyses the root cause of certain problems as they arise to allow the company to learn from past situations and inform improvements to our products, policies, processes, or enforcement operations.

Research and expert consultation also play a major role in Meta's product development process. For example, over the past few years, Meta has refined how our teams design products for young people based on external guidance from organisations like the UN, the OECD and children's rights groups. Through our global co-design programme, we also regularly consult with third party experts and young people, parents and guardians to make sure we build products that meet their needs. We developed a process to help us apply the UN's Convention on the Rights of the Child directly to the products and experiences we build at Meta. We complemented our own internal research with input from global data protection regulators to create Meta's Best Interests of the Child Framework (available here: https://www.ttclabs.net/news/metas-best-interests-of-the-child-framework), which distils the "best interests of the child" standard into six key considerations that product teams can consult throughout the development process. The framework is available as a resource for all employees at Meta and is intentionally applied at different points of the product development cycle. Each consideration has extensive guiding questions, resources and examples to help our teams and product builders make balanced decisions.

We also have a number of global Risk Ops teams that focus on risks relating to user and platform safety:

- Risk Prevention team – works to mitigate known gaps in scaled detection and enforcement of content risks.
- Risk Management & Intelligence team – works with various other teams to identify and provide mitigations, and identify ownership for any systemic problems unearthed.
- Risk Intelligence team – works with relevant teams to conduct incident reviews of escalations that take place, as well as root cause analysis.
- Imminent Risk team – focuses on geopolitical issues and related activity on our platforms, and provides stop gap mitigations in advance of product/policy intervention.
- Integrity Product Operations Centre (IPOC) – the forum through which crisis response is coordinated.

As an example of how features of a product can affect the risks posed by violating content, the following features exacerbate risks posed by violating content on Dating (a feature within the Facebook App):

- Dating is a product that serves the purpose of helping users find romantic connections, which exacerbates the risks of users sharing unwanted sexual imagery.
- Dating is a product whose users are generally aiming to build relationships and trust in private one-on-one conversations with other users they believe are doing the same. Users may therefore have extra motivation to believe and interact with other users they match with, which exacerbates the risk of fraud and scammers using the Dating product.

The following features mitigate risks posed by violating content on the Dating product:

- When users initiate a match, a message is displayed in their chat thread informing them that they should report anyone who asks for money or sensitive information, with a link to further Safety Tips.
- Dating users can block someone on Dating, either pre-emptively or based on interactions with that user on Dating, to mitigate risks of encountering violating content from that person (e.g. public order offences, harassment, or stalking).
- Dating users cannot send photos in the Conversations (chat) feature, which limits the risks of sharing of violating content such as unwanted sexual images.
- Dating users can specifically report another user based on a series of categories including, but not limited to, sharing inappropriate things (e.g., violence, nudity & sexual services, drugs, guns, sexual messages), the user being under 18, engaged in a scam, harassment, or use of the profile for a business or self promotion.

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

Meta has a range of teams and structures that contribute to user and platform safety, which are overseen by Meta's management.

For example, Meta employs over 40,000 people to work on safety and security, and these teams are spread across almost every vertical function of our company (including Engineering, Product Management, Operations, Business, Policy, and Legal functions).

At board level, the Audit and Risk Oversight Committee ("AROC") of the Meta Board, to which the board of Meta Platforms Ireland Limited reports, has several responsibilities, including oversight of the company's risks related to social responsibility. As set out in its Charter, AROC reviews the company's assessment of the major ways in which our services can be used to facilitate harm or undermine public safety or the public interest, as well as the steps the company has taken to monitor, mitigate, and try to prevent such abuse. AROC conducts these reviews together with management at least annually, and generally are briefed twice a year by the VP, Integrity and the VP, Content Policy on Community Safety and Security issues.

We also have a Whistleblower and Complaint Policy, and a channel by which employees can raise any concern, including in relation to products or safety. This Policy and channel are highlighted in our Code of Conduct, internal training courses, and internal communications tools. There are other less formal methods by which employees may raise concerns, including through 'office hours' with various teams, or directly with managers, human resources business partners, and our employment legal and compliance teams.

## TERMS OF SERVICE AND POLICY STATEMENTS

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

There is no singular way to ensure clarity and accessibility of terms of service and policy statements. Criteria that Meta considers to be relevant include, but are not limited to, the following:

- Terms and policies should be available prior to registration, and when users are not logged into the service;
- Users should be able to access terms and policies easily;
- Terms and policies should be laid out in a way that makes it easy to locate information;
- Terms and policies should use clear and simple language; and
- There should be regular and timely reminders of terms and policies, for example when content is reported or reports are processed.

Meta takes a range of steps to help make the terms for use of its services and other policy statements easily available and accessible. We offer tools to help people make safe choices on our platforms and we work to be transparent about how we address these issues.

That's why we make our policies – including the Facebook Terms of Service and Community Standards, the Instagram Terms of Use and Community Guidelines, and other specific policies (such as our Pages Groups and Events Policy, Commerce Policies and Ad Policies) – available online to everyone, as linked below, to ensure that they are easy to locate:

- Facebook Terms of Service: https://www.facebook.com/terms.php
- Facebook Community Standards: https://www.facebook.com/communitystandards
- Facebook Pages Groups and Events Policy: https://business.facebook.com/policies/pages_groups_events/
- Facebook Ads Policies: https://www.facebook.com/policies/ads/
- Instagram Terms of Use: https://help.instagram.com/581066165581870
- Instagram Community Guidelines: https://help.instagram.com/477434105621119/
- Meta Commerce Policies: https://www.facebook.com/policies_center/commerce/

We also make certain policy statements available via Meta Newsroom: https://about.fb.com/news/

Our terms and policies are written clearly and straightforwardly, and are made available in a range of languages, to make them easy to understand. Applicable terms are brought to users' attention during the sign-up process, and we share the changes we make to our policies each month. In addition, if a user posts something that goes against our policies, Meta may provide a notice to the user which may contain reference to the relevant part of the policies and a description of why Meta doesn't allow the content.

Meta also releases a quarterly CSER, which shows how Meta is doing at enforcing its policies. This kind of transparency lets people see clearly how we are addressing safety issues and helps us get much-needed feedback. Our CSER are publicly available here:

https://transparency.fb.com/data/community-standards-enforcement/.

In addition, Meta established the Data Transparency Advisory Group (DTAG) in 2018. This is an independent body made up of international experts in measurement, statistics, criminology and governance. Their task was to provide an independent, public assessment of whether the metrics we share in the CSER provide accurate and meaningful measures of Facebook's content moderation challenges and our work to address them. In its assessment, the advisory group notes the CSER is an important exercise in transparency. But they also highlight other areas where we could be clearer in order to build more awareness and responsiveness to the people who use our platform. This includes recommendations of ways to enhance community input into our governance model and concrete suggestions on how to expand the information provided in the report to give people more context. For a summary of the DTAG's findings see here: https://about.fb.com/news/2019/05/dtag-report/ and https://research.facebook.com/blog/2019/5/exploring-feedback-from-data-and-governance-experts-a-research-based-response-to-the-data-transparency-advisory-group-report/

Meta is also a founding member of the Digital Trust and Safety Partnership, which has brought together internet platforms to agree on high level best practices for content moderation and is working to add more detail and specifics, which will vary depending on the type of services. This framework is the first that details what platforms need to do to ensure trust and safety for their services, in a way that is both platform- and content-agnostic and serves as a model for assessment of trust and safety on any digital service. In addition to the DTSP Best Practices, the organisation has also created a self-assessment framework (the "Safe Framework") that companies can use to measure their adherence to DTSP Best Practice. Over the first quarter of 2022, DTSP member companies conducted self-assessments, focused on either a product or problem area, using the Safe Framework. DTSP published a new report entitled "The Safe Assessments: An Inaugural Evaluation of Trust & Safety Best Practices" (available here: https://dtspartnership.org/dtsp-safe-assessments-report/). The report synthesises self-assessment submissions from ten DTSP members and provides an anonymised snapshot of industry's posture regarding digital trust and safety. This report reflects DTSP's first use of the Safe Framework in practice. We are encouraged by the results, and know that our learnings from this initial effort will help us to iterate and make the Best Practices framework more effective over time.

**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?**

Please see our response to question 2 above for an overview of how our standards apply to content that is also illegal in the UK and address other content that is objectionable or harmful. To provide a more detailed example, we set out below the types of content prohibited on Facebook, in line with the terms and policies referenced above.

Development and Maintenance of Policies

We take great care to craft policies that are inclusive of different views and beliefs – in particular, those of people and communities that might otherwise be overlooked or marginalised.

In terms of maintaining our policies, there are a number of reasons why Meta may draft a new policy or revise an existing one:

- Meta's Content Policy team, which sits in more than a dozen locations around the world, is responsible for developing our Community Standards and Community Guidelines. The team includes subject matter experts on issues such as hate speech, child safety and terrorism, as well as people with experience in criminal prosecution, rape crisis counselling, academics, human rights, law and education. Many have also worked on issues of voice and safety long before coming to Meta.

- The Integrity team assesses the global impact of potential policy changes and builds the technology to scale the detection and enforcement of new policies.
- The Global Operations team, whose employees, contractors and outsourcing partners are responsible for enforcing our policies, keep us informed about trends or times when we may need to clarify a policy.
- Research teams may also point us to data or user sentiment that seems best addressed through policymaking.
- Sometimes, we identify a point in our policy that could be supplemented, or an external stakeholder tells us that a policy fails to adequately address an issue that's important to them. In other cases, the press draws attention to a point that requires supplementation.

Facebook's and Instagram's policies area also reviewed by civil society organisations, activist groups, and thought leaders, in such areas as digital and civil rights, anti-discrimination, free speech, and human rights. We also engage with academics who have relevant expertise. Academics may not directly represent the interests of others, but they are important stakeholders by virtue of their extensive knowledge, which helps us create better policies for everyone.

Our work with experts includes work with the Safety Advisory Board (see here for more information: https://www.facebook.com/help/222332597793306/?ref=sc), and we gather feedback from our community to develop policies, tools and resources to help keep people safe. The Safety Advisory Board is composed of leading internet safety organisations from around the world, including Childnet International, National Network to End Domestic Violence, ConnectSafely, FOSI, Net Family News, Centre for Social Research and Telefono Azzurro. We have built a network of well over 500 safety organisations covering topics from child safety to suicide prevention to LGBTQ advocacy in order to ensure we are getting diverse, high-quality advice and input.

Moreover, in 2019 we made the decision to set up a regular check in with experts from over 20 countries to discuss some of the complex issues associated with suicide and self-injury content, revisit decisions we have made to ensure they align with the latest research, and ensure we are doing our best to support all those on our platform. We cover a wide range of issues during these discussions, including how should we deal with suicide notes posted on our platform, what are the risks associated with viewing aggregated sad content online, and when should we allow newsworthy depictions of suicide. We also seek their input on product enhancements to foster the well-being of our community. Further information is available here:

https://www.facebook.com/safety/wellbeing/suicideprevention/expertengagement.

In addition, a meeting called the Policy Forum takes place on a regular basis where we discuss potential changes to our Community Standards, Community Guidelines, Advertising Policies or Product Policies. At this meeting, subject matter experts from the Content Policy team propose adding new policies or amending existing ones. These meetings help the team factor in cultural differences on what is acceptable and better understand broad perspectives on safety and voice and the impact of our policies on communities globally.

A variety of internal stakeholders also participate in the meetings. This includes team members from safety and cybersecurity policy, Global Operations, Civil Rights and Human Rights, legal, communications and diversity, as well as counterterrorism specialists, product managers and other public policy leads.

Our Content Policy team typically gives two types of presentations at the meetings: a heads-up or a policy recommendation. A heads-up is a short presentation that introduces an issue the team plans to work through, with internal and external input. After the team has received input, analysed relevant data about the issue and prepared options for updating a policy, subject matter experts will present a recommendation so the larger group can discuss it.

Our policies evolve over time based on feedback from these meetings, as well as changes in social norms, language and product updates.

Following the Policy Forum meeting, the policy changes are prepared for implementation, which takes time, and actual launch dates vary. Once the policy has been launched, we publish meeting minutes from the Policy Forum and note these changes in our Community Standards, viewable in the Change log (also available here: https://transparency.fb.com/en-gb/policies/improving/policy-forum-minutes).

Types of content prohibited on Facebook and Instagram

The relevant Facebook terms and policies are available via the following links:

- Facebook Terms of Service: https://www.facebook.com/terms.php
- Facebook Community Standards: https://www.facebook.com/communitystandards
- Facebook Pages Groups and Events Policy: https://business.facebook.com/policies/pages_groups_events/
- Meta Commerce Policies: https://www.facebook.com/policies_center/commerce/

The relevant Instagram terms and policies are available via the following links:

- Instagram Terms of Use: https://help.instagram.com/581066165581870
- Instagram Community Guidelines: https://help.instagram.com/477434105621119/

Section 3.2 of the Facebook Terms of Service and the Instagram Terms of Use include a blanket prohibition on using our products to do or share anything that is unlawful.

Below are examples of the specific types of content that the terms and policies above prohibit:

*Violence and Incitement*

We aim to deter potential offline harm that may be related to content on Facebook. While we understand that people commonly express disdain or disagreement by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence and in many cases will also remove threats of less severe violence. We remove content, disable accounts, and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety. We also try to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety. In determining whether a threat is credible, we may also consider additional information like a person's public visibility and the risks to their physical safety.

In some cases, we see aspirational or conditional threats directed at terrorists and other violent actors (e.g. “Terrorists deserve to be killed”), and we deem those non-credible absent specific evidence to the contrary.

*Dangerous Individuals and Organisations*

In an effort to prevent and disrupt real-world harm, we do not allow any organisations or individuals that proclaim a violent mission or are engaged in violence to have a presence on Facebook. This includes organisations or individuals involved in the following:

- Terrorist activity
- Organised hate
- Mass murder (including attempts) or multiple murder
- Human trafficking
- Organised violence or criminal activity

We also remove content we find or learn about that expresses support or praise for groups, leaders, or individuals involved in these activities.

*Coordinating Harm and Publicising Crime*

In an effort to deter and disrupt offline harm and copycat behaviour, we prohibit people from facilitating, organising, promoting, or admitting to certain criminal or harmful activities targeted at people, businesses, property or animals. We allow people to debate and advocate for the legality of criminal and harmful activities, as well as draw attention to harmful or criminal activity that they may witness or experience as long as they do not advocate for or coordinate harm.

*Restricted Goods & Services*

To encourage safety and compliance with common legal restrictions, we prohibit attempts by individuals, manufacturers, and retailers to purchase, sell, or trade non-medical drugs, pharmaceutical drugs, and marijuana. We also prohibit the purchase, sale, gifting, exchange, and transfer of firearms, including firearm parts or ammunition, between private individuals on Facebook. Some of these items are not regulated everywhere; however, because of the borderless nature of our community, we try to enforce our policies as consistently as possible. Firearm stores and online retailers may promote items available for sale off of our services as long as those retailers comply with all applicable laws and regulations. We allow discussions about sales of firearms and firearm parts in stores or by online retailers and advocating for changes to firearm regulation. Regulated goods that are not prohibited by our Community Standards may be subject to our more stringent Commerce Policies.

*Fraud and Deception*

In an effort to deter and disrupt harmful or fraudulent activity, we remove content that we find or learn about aimed at deliberately deceiving people to gain an unfair advantage or deprive another of money, property, or legal right. However, we allow people to raise awareness and educate others as well as condemn these activities using our platform.

*Suicide and Self-Injury*

In an effort to promote a safe environment on Facebook, we work to remove content that we find or learn about that encourages suicide or self-injury, including certain graphic imagery and real-time depictions that experts tell us might lead others to engage in similar behaviour. We also remove any content that we find or learn about that identifies and negatively targets victims or survivors of self-injury or suicide seriously, humorously, or rhetorically.

Self-injury is defined as the intentional and direct injuring of the body, including self-mutilation and eating disorders. We want Facebook to be a space where people can share their experiences, raise awareness about these issues, and seek support from one another, which is why we allow people to discuss suicide and self-injury without encouraging it.

We work with organisations around the world to provide assistance to people in distress. We also talk to experts in suicide and self-injury to help inform our policies and enforcement. For example, we have been advised by experts that we should not remove live videos of self-injury while there is an opportunity for loved ones and authorities to provide help or resources.

*Child Sexual Exploitation, Abuse and Nudity*

We do not allow content that sexually exploits or endangers children. When we become aware of apparent child exploitation, we report it to the National Center for Missing and Exploited Children (NCMEC), in compliance with applicable law. We know that sometimes people share nude images of their own children with good intentions; however, we generally remove these images because of the potential for abuse by others and to help avoid the possibility of other people reusing or misappropriating the images.

We also work with external experts, including the Facebook Safety Advisory Board, to discuss and improve our policies and enforcement around online safety issues, especially with regard to children.

*Adult Sexual Exploitation*

We recognise the importance of Facebook as a place to discuss and draw attention to sexual violence and exploitation. We believe this is an important part of building common understanding and community. In an effort to create space for this conversation while promoting a safe environment, we remove content that depicts, threatens or promotes sexual violence, sexual assault, or sexual exploitation, while also allowing space for victims to share their experiences. We remove content that displays, advocates for, or coordinates sexual acts with non-consenting parties or commercial sexual services, such as prostitution and escort services. We do this to avoid facilitating transactions that may involve trafficking, coercion, and non-consensual sexual acts.

To protect victims and survivors, we also remove images that depict incidents of sexual violence and intimate images shared without permission from the people pictured. We’ve written about the technology we use to protect against intimate images and the research that has informed our work. We’ve also put together a guide to reporting and removing intimate images shared without consent.

*Bullying and Harassment*

Bullying and harassment happen in many places and come in many different forms, from making threats to releasing personally identifiable information, to sending threatening messages, and making unwanted malicious contact. We do not tolerate this kind of behaviour because it prevents people from feeling safe and respected on Facebook.

We distinguish between public figures and private individuals because we want to allow discussion, which often includes critical commentary of people who are featured in the news or who have a large public audience. For public figures, we remove attacks that are severe as well as certain attacks where the public figure is directly tagged in the post or comment. For private individuals, our protection goes further: we remove content that's meant to degrade or shame, including, for example, claims about someone's sexual activity. We recognise that bullying and harassment can have more of an emotional impact on minors, which is why our policies provide heightened protection for users between the ages of 13 and 18.

Context and intent matter, and we allow people to share and re-share posts if it is clear that something was shared in order to condemn or draw attention to bullying and harassment. In certain instances, we require self-reporting because it helps us understand that the person targeted feels bullied or harassed. In addition to reporting such behaviour and content, we encourage people to use tools available on Facebook to help protect against it.

We also have a Bullying Prevention Hub, which is a resource for teens, parents, and educators seeking support for issues related to bullying and other conflicts. It offers step-by-step guidance, including information on how to start important conversations about bullying.

*Human Exploitation*

In an effort to disrupt and deter harm, we remove content that facilitates or coordinates the exploitation of humans, including human trafficking. We define human trafficking as the business of depriving someone of liberty for profit. It is the exploitation of humans in order to force them to engagein commercial sex, labor, or other activities against their will. It relies on deception, force and coercion, and degrades humans by depriving them of their freedom while economically or materially benefiting others.

Human trafficking is multi-faceted and global; it can affect anyone regardless of age, socioeconomic background, ethnicity, gender, or location. It takes many forms, and any given trafficking situation can involve various stages of development. By the coercive nature of this abuse, victims cannot consent.

While we need to be careful not to conflate human trafficking and smuggling, the two can be related and exhibit overlap. The United Nations defines human smuggling as the procurement or facilitation of illegal entry into a state across international borders. Without necessity for coercion or force, it may still result in the exploitation of vulnerable individuals who are trying to leave their country of origin, often in pursuit of a better life. Human smuggling is a crime against a state, relying on movement, and human trafficking is a crime against a person, relying on exploitation.

*Privacy Violations and Image Privacy Rights*

Privacy and the protection of personal data are fundamentally important values for Meta. We work hard to safeguard people's personal identity and information, and we do not allow people to post personal or confidential information about themselves or others. We also provide people ways to report imagery that they believe to be in violation of their privacy rights.

Internally, privacy and data protection are everyone's responsibility at Meta – from our CEO and executives, to engineers and sales teams across the globe, and everyone across the company, we are all responsible for privacy. As a result, we have a cross-functional group of stakeholders across the company who provide engineering, legal, policy, compliance, and product expertise that enable the design and implementation of our privacy program.

*Hate Speech*

We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence.

We define hate speech as a direct attack on people based on what we call protected characteristics and classify the below list as "protected groups". We define attack as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation.

"Protected groups" under our policies and Community Standards:

- Race
- Ethnicity
- National origin and nationality
- Religious affiliation
- Caste
- Sex
- Gender
- Gender identity
- Sexual orientation or practices
- Disability
- Medical condition
- Migrants
- Refugees
- Immigrants
- Asylum seekers
- Age (when paired with another protected characteristic)

Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others. In some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. People sometimes express contempt in the context of a romantic break-up. Other times, they use gender-exclusive language to control membership in a health or positive support group, such as a breastfeeding group for women only. In all of these cases, we allow the content but expect people to clearly indicate their intent, which helps us better understand why they shared it. Where the intention is unclear, we may remove the content.

*Violent and Graphic Content*

We remove content that glorifies violence or celebrates the suffering or humiliation of others because it may create an environment that discourages participation. We allow graphic content (with some limitations) to help people raise awareness about issues. We know that people value the ability to discuss important issues like human rights abuses or acts of terrorism. We also know that people have different sensitivities with regard to graphic and violent content. For that reason, we add a warning label to especially graphic or violent content so that it is not available to people under the age of eighteen and so that people are aware of the graphic or violent nature before they click to see it. In some cases, such as where imagery of police-inflicted violence is shared, we add a warning screen but do not limit distribution to people under the age of eighteen, to further facilitate awareness-raising.

*Adult Nudity and Sexual Activity*

We restrict the display of nudity or sexual activity because some people in our community may be sensitive to this type of content. Additionally, we default to removing sexual imagery to prevent the sharing of non-consensual or underage content. Restrictions on the display of sexual activity also apply to digitally created content unless it is posted for educational, humorous, or satirical purposes.

Our nudity policies have become more nuanced over time. We understand that nudity can be shared for a variety of reasons, including as a form of protest, to raise awareness about a cause, or for educational or medical reasons. Where such intent is clear, we make allowances for the content. For example, while we restrict some images of female breasts that include the nipple, we allow other images, including those depicting acts of protest, women actively engaged in breast-feeding, and photos of post-mastectomy scarring. We also allow photographs of paintings, sculptures, and other art that depicts nude figures.

*Adult Sexual Solicitation and Sexually Explicit Language*

As noted in Section 8 of our Community Standards (Adult Sexual Exploitation), people use Facebook to discuss and draw attention to sexual violence and exploitation. We recognise the importance of and want to allow for this discussion. We draw the line, however, when content facilitates, encourages or coordinates sexual encounters between adults. We also restrict sexually explicit language that may lead to solicitation because some audiences within our global community may be sensitive to this type of content and it may impede the ability for people to connect with their friends and the broader community. In contrast, we allow for the discussion of sex worker rights advocacy and changes to sex work regulation.

## REPORTING AND COMPLAINTS

### 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?

Further to our response to question 5 above, Meta takes a range of steps to improve transparency, accessibility, ease of use and awareness of its product features. We provide more information in the following questions about our reporting and complaints mechanisms.

### 8. If your service has <u>reporting or flagging</u> mechanisms in place for illegal content, or users who post illegal content, how are these processes designed and maintained?

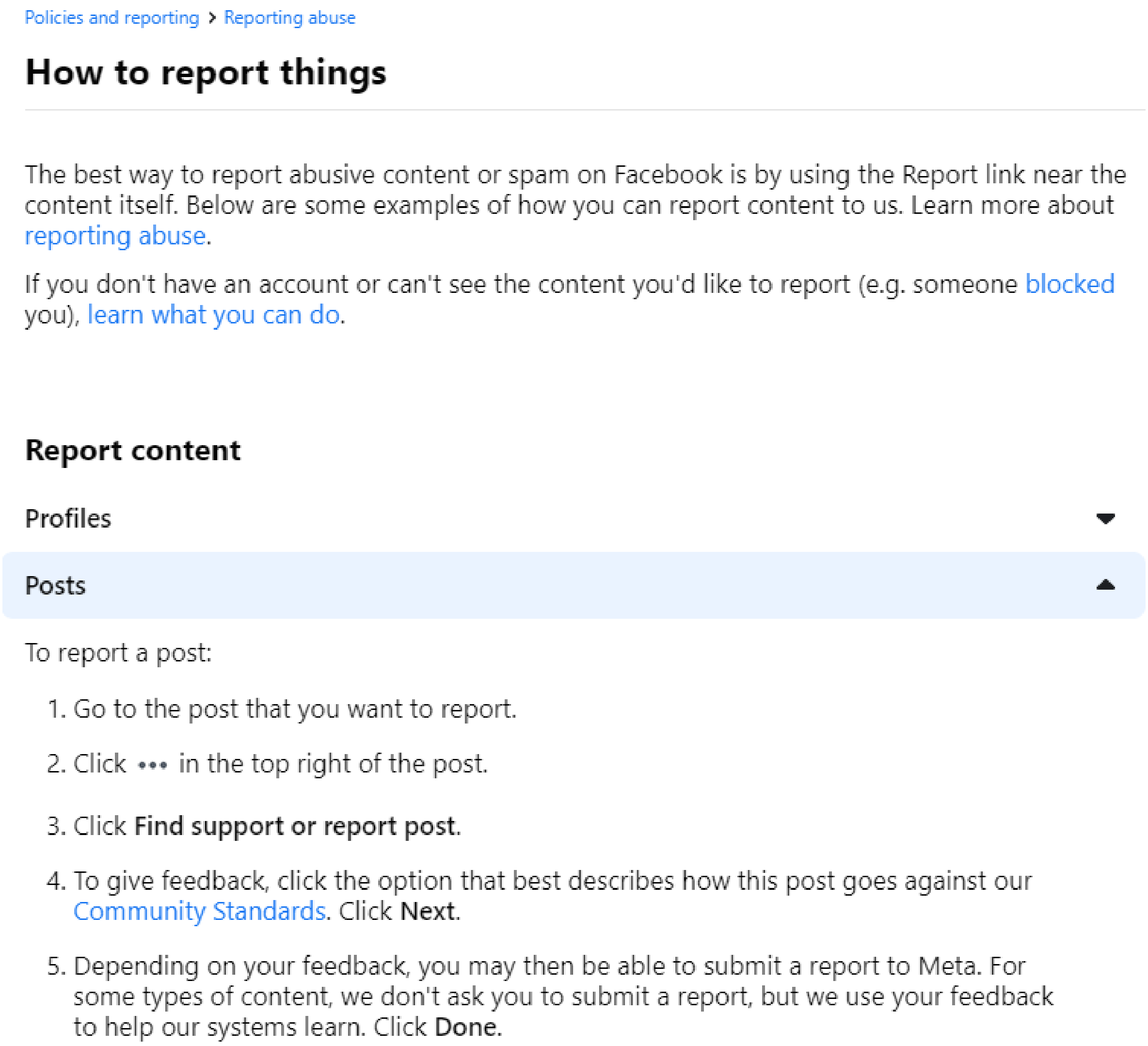
Meta maintains various reporting / flagging mechanisms across its services.

<u>Standards Reporting</u>

Content on Facebook and Instagram may be reported where it is prohibited by the Facebook Community Standards or Instagram Community Guidelines, which outline what is and is not allowed on the Services and cover various categories of prohibited content. As noted above, a category may cover illegal content under UK law and / or legal content that is identified as being harmful or inappropriate, e.g. content that has been identified by different regulations or experts as particularly harmful for young people, such as content depicting nudity and sexual activity, sexual solicitation, violent and graphic content, and hate speech. A Community Standards report does not trigger a legal review by us – in accordance with the user's Community Standards report, we review for violation of our Community Standards. We maintain separate reporting mechanisms for users to report content they believe violates the local law (described in further detail below).

On Facebook and Instagram, standards reporting is available to logged-in users through the '3 dots' feature. These are easy to access and use, even for the youngest user. Specific reporting steps and the information provided depend on content type and surface, which are explained in the Facebook and Instagram Help Centres (linked below, with screenshots of the reporting steps for posts on Facebook and the Instagram app):

https://www.facebook.com/help/1380418588640631



Policies and reporting > Reporting abuse

## How to report things

The best way to report abusive content or spam on Facebook is by using the Report link near the content itself. Below are some examples of how you can report content to us. Learn more about reporting abuse.

If you don't have an account or can't see the content you'd like to report (e.g. someone blocked you), learn what you can do.

### Report content

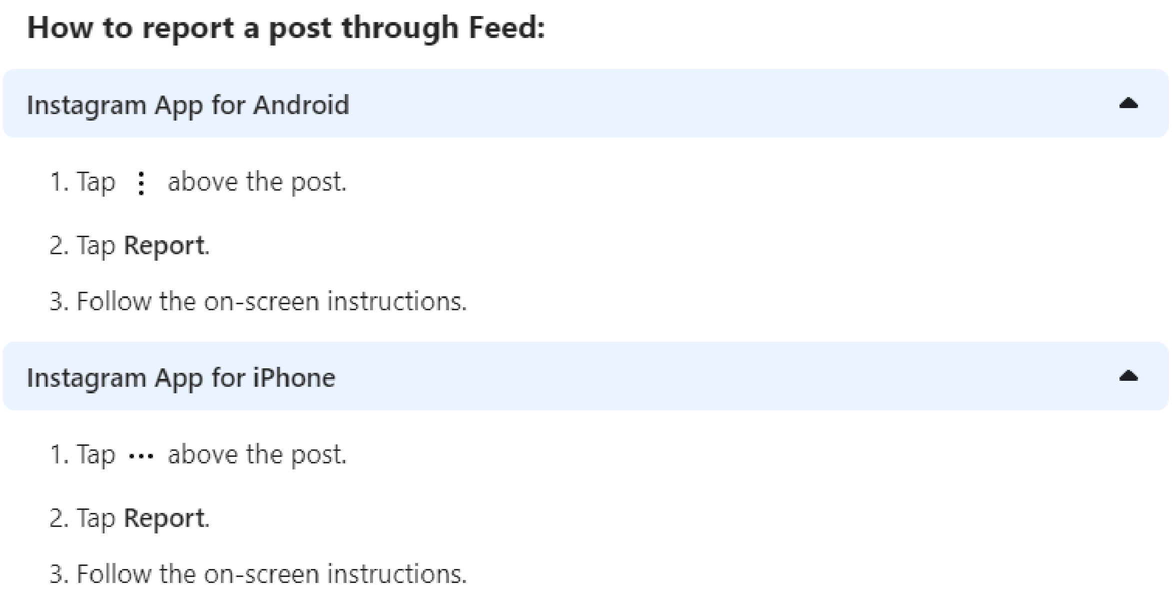
Profiles ▼

Posts ▲

To report a post:

1. Go to the post that you want to report.
2. Click ••• in the top right of the post.
3. Click **Find support or report post**.
4. To give feedback, click the option that best describes how this post goes against our Community Standards. Click **Next**.
5. Depending on your feedback, you may then be able to submit a report to Meta. For some types of content, we don't ask you to submit a report, but we use your feedback to help our systems learn. Click **Done**.

https://help.instagram.com/192435014247952



**How to report a post through Feed:**

Instagram App for Android ▲

1. Tap ⋮ above the post.
2. Tap **Report**.
3. Follow the on-screen instructions.

Instagram App for iPhone ▲

1. Tap ⋯ above the post.
2. Tap **Report**.
3. Follow the on-screen instructions.

In addition, Marketplace and Dating users can report violations that are specific to these products. For example, on Marketplace, users can report buyers, sellers, and items for violations of our policies.

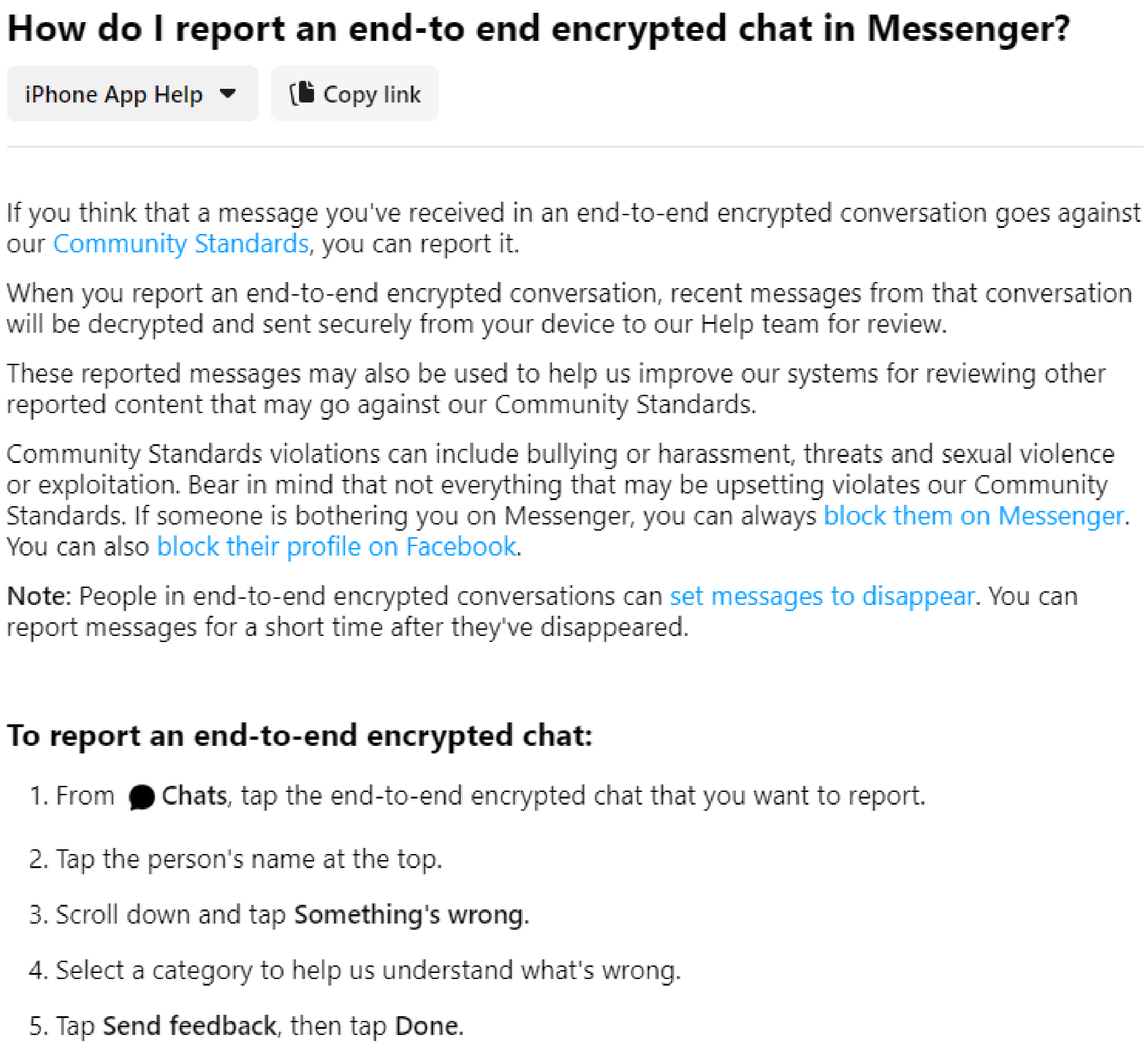
On Dating, users can access reporting/flagging if they have created a Dating account. Dating users can report another user's profile, media (e.g. photos), conversations/messages, and Stories on their profile. In addition to violating content, users can also report various types of violating conduct, including but not limited to:

- Use of a fake account
- Sharing inappropriate content (e.g. violence, nudity & sexual services, drugs, guns, sexual messages)
- Profile under the age of 18
- Scam activity (e.g. romance scam, solicitation, asking for money, asking for personal information)
- Harassment (e.g. hate speech, unwanted messages, offline harassment)
- An account that appears to be a business
- Conduct relating to suicide or self-injury
- Other types of violating conduct, including self promotion and spam activity

The report types made available to users are based on prevalence on the platform, which has been verified through research and survey of users and the logging of report types over time. Moreover, particular consideration has been given to the needs of different user groups. For example, non-cisgender users (e.g. trans users) are at a higher risk of facing harm from content, particularly threats to kill or fear of violence, on the Dating platform. These users are given controls over who they are suggested to (for example, people looking to date everyone vs. trans men specifically) as a mitigation to protect themselves from violence due to prejudice against their gender identity. Conversely, given that the Dating product is not to be used by users who are under 18 years old, it does not include reporting mechanisms that have been specifically designed for use by children.

Following our review of a report, the person making the report will receive a notification informing them of the outcome of their report.

Standards reporting is also available on Messenger, including for end-to-end encrypted conversations on the Messenger app. As an example, see below a screenshot and link to the specific reporting steps for end-to-end encrypted conversations, as set out in the Messenger Help Centre (in this example, for the Messenger app on iPhone).



## How do I report an end-to end encrypted chat in Messenger?

If you think that a message you've received in an end-to-end encrypted conversation goes against our Community Standards, you can report it.

When you report an end-to-end encrypted conversation, recent messages from that conversation will be decrypted and sent securely from your device to our Help team for review.

These reported messages may also be used to help us improve our systems for reviewing other reported content that may go against our Community Standards.

Community Standards violations can include bullying or harassment, threats and sexual violence or exploitation. Bear in mind that not everything that may be upsetting violates our Community Standards. If someone is bothering you on Messenger, you can always block them on Messenger. You can also block their profile on Facebook.

**Note**: People in end-to-end encrypted conversations can set messages to disappear. You can report messages for a short time after they've disappeared.

### To report an end-to-end encrypted chat:

1. From **Chats**, tap the end-to-end encrypted chat that you want to report.
2. Tap the person's name at the top.
3. Scroll down and tap **Something's wrong**.
4. Select a category to help us understand what's wrong.
5. Tap **Send feedback**, then tap **Done**.

https://www.facebook.com/help/messenger-app/498828660322839/

Legal Removal Requests

Meta also provides a legal removal request form, which allows individuals in the UK, including both logged-in and non-logged-in users, to report content they believe violates their personal legal rights or applicable local laws. This form allows people in the UK to report alleged violations of intellectual property, defamation, right to privacy / erasure, and violations of other laws not covered by the first three categories. The form is accessible via the Facebook and Instagram Help Centres:

https://www.facebook.com/help/contact/319149701968527 https://help.instagram.com/contact/406206379945942

Trusted Partner programme

The Trusted Partner programme is a critical part of Facebook's efforts to keep our platforms safe. Trusted Partners help us detect problematic content, analyse harmful content trends, and enhance the contextual understanding needed to enforce our Community Standards.

The goal of the Trusted Partner programme is to establish and maintain relationships and reporting channels with expert organisations that raise queries and concerns about content on Facebook and Instagram, in order to:

- address problematic content trends and prevent harm,
- foster online safety and security,
- increase transparency, and
- improve our content policies.

Our network of Trusted Partners includes over 400 non-governmental organisations, humanitarian agencies, human rights defenders and researchers from 77 countries around the globe. Our network of partners helps Facebook and Instagram to:

- learn from local experts across the globe,
- identify and address policy, process, or training gaps to improve the standard in-product report/review process, and
- ensure that our review of high-priority content and accounts is informed by critical and up-to-date context.

The Trusted Partner Channel (TPC) provides our partner organisations with an expedited mechanism for reporting content to Facebook. This channel is distinct from standard in-product reporting (which is open to all users and which routes content to our scaled review centres around the globe). The TPC, by contrast, is staffed by escalations specialists who triage reports and route them to expert teams for analysis, including in-depth investigations where appropriate. In contrast to standard reporting where users select a series of options in requesting review of problematic content, Trusted Partners are able to provide additional context, which enables deeper investigation and analysis of the content in question.

The same content policies are enforced in both the TPC and standard in-app reporting. Across reporting channels, we prioritise reports related to imminent harm and will only remove the content and accounts that are in violation of our Community Standards.

### 9. If your service has a complaints mechanism in place, how are these processes designed and maintained?

Please see our response to question 8 above in relation to reporting / flagging mechanisms available on our services. Where we action reported content, Facebook and Instagram users generally have access to an appeals process.

In most cases, if someone publishes a post which we decide to remove from Facebook or Instagram for going against our policies, the person who posted it is notified, and given the option to accept the decision or disagree and request another review. If they choose to disagree with the decision, the content is resubmitted for another review. The content is not visible to other people on Facebook or Instagram while we review it again.

If the reviewer accepts the original decision, the content remains off Facebook or Instagram. However, if the reviewer disagrees with the initial review and decides it should not have been removed, the content will go to another reviewer. This reviewer's decision will determine whether the content should be on Facebook or Instagram or not.

Today, we offer appeals for the vast majority of violation types on Facebook and Instagram. We do not offer appeals for violations with extreme safety concerns, such as child exploitation imagery. We use a combination of human review and technology to process appeal requests. During busy periods, we may not always be able to review everything based on our review capacity. In addition, given Covid-19 related limitations on our content moderation systems, we have had to turn off appeals in some areas. We are working to enable more content review resources online and turn these appeals back on. In the meantime, users have the opportunity to tell us when they disagree with a content moderation decision we have made, and we're monitoring that feedback to improve our accuracy.

We also provide appeals not just for content that we took action on, but also for content that was reported but not acted on.

Our CSER also includes more information about content appeals and content restored following appeal. See 'Appealed content' (https://transparency.fb.com/lt-lt/policies/improving/appealed-content-metric/) and 'Restored content' (https://transparency.fb.com/lt-lt/policies/improving/restored-content-metric/).

### 10. What action does your service take in response to reports or complaints?

Taking action on content could include removing a piece of content from Facebook or Instagram, covering photos or videos that may be disturbing to some audiences with a warning, or disabling accounts.

It is important to note that some of the most commonly asked questions about user reports do not in fact provide meaningful insights into the efficacy of a platform's integrity efforts. This is due to a number of common assumptions about user reports which unfortunately are not correct. These include:

1. That user reports are a good measure of violation of policy or legality.
2. That user reports attribution (report reason) is accurate.
3. That we review all user reports.
4. That we optimise for review time when it comes to user reports.

**None of these assumptions are true.**

A more meaningful metric would be in relation to action on confirmed violations, instead of the massive volume that is reported but not violating. Asking about the proportion of user reports that are actioned is more a reflection of the quality of users' reporting, and not our ability to manage violating content.

Further, in recent years, technology has started to play a more central role in our content enforcement operations. Our algorithms are getting better all the time at identifying content that obviously violates our Community Standards and automatically taking it down before anyone sees it. For example, in our latest CSER we shared that in Q2 of 2022 we found and removed 99.5% of the violating content we actioned for Violent and Graphic content before people reported it.

By way of example, we set out below some metrics from our latest CSER regarding actioned content on Facebook and Instagram. The numbers below represent actions taken on content in Q2 2022.

Facebook:

- Adult Nudity and Sexual Activity: 38.4 million pieces of content actioned, 97.2% actioned before being reported by users.
- Bullying and Harassment: 8.2 million pieces of content actioned, 76.8% of content actioned before being reported by users.
- Child Nudity and Physical Abuse: 1.9 million pieces of content actioned, 97.3% of content actioned before being reported by users.
- Child Sexual Exploitation: 20.4 million pieces of content actioned, 99.1% of content actioned before being reported by users.
- Dangerous Organisations: Terrorism and Organised Hate: 13.6 million pieces of terrorism content actioned and 2.3 million pieces of organised hate content actioned, 98.9% of terrorism content and 96.9% of organised hate content actioned before being reported by users.
- Hate Speech: 13.5 million pieces of content actioned, 95.6% of content actioned before being reported by users.
- Regulated Goods: Drugs and Firearms: 3.9 million pieces of drug content actioned and 1.6 million pieces of firearms content actioned, 98.1% of drug content and 94.4% of firearms content actioned before being reported by users.
- Suicide and Self-Injury: 11.3 million pieces of content actioned, 99.1% of content actioned before being reported by users.
- Violent and Graphic Content: 45.9 million pieces of content actioned, 99.5% of content actioned before being reported by users.

Instagram:

- Adult Nudity and Sexual Activity: 10.3 million pieces of content actioned, 94.3% of content actioned before being reported by users.
- Bullying and Harassment: 6.1 million pieces of content actioned, 87.4% of content actioned before being reported by users.
- Child Nudity and Physical Abuse: 480,000 pieces of content actioned, 93.4% of content actioned before being reported by users.
- Child Sexual Exploitation: 1.2 million pieces of content actioned, 94.9% of content actioned before being reported by users.
- Dangerous Organisations: Terrorism and Organised Hate: 1.9 million pieces of terrorism content actioned and 449,000 pieces of organised hate content actioned, 93.3% of terrorism content and 87.6% of organised hate content actioned before being reported by users.
- Hate Speech: 3.8 million pieces of content actioned, 91.2% of content actioned before being reported by users.
- Regulated Goods: Drugs and Firearms: 1.9 million pieces of drug content actioned and 215,000 pieces of firearms content actioned, 96.8% of drug content and 93.6% of firearms content actioned before being reported by users.
- Suicide and Self-Injury: 6.4 million pieces of content actioned, 98.4% of content actioned before being reported by users.
- Violent and Graphic Content: 10.2 million pieces of content actioned, 99.3% of content actioned before being reported by users.

Our approach to content moderation and prioritisation is addressed in our newsroom post here (https://about.fb.com/news/2020/08/how-we-review-content/), but simply put, content is ranked according to multiple factors such as virality, severity of harm and likelihood of violation. In an instance where the system is near-certain that content is breaking the Community Standards, it may remove it. Where there is less certainty, it will prioritise the content for teams to review.

As noted in question 9, most of the content disables and content takedowns are appealable on Facebook and Instagram. The user receives a notice in feed and allows the user to continue through an explanation as to how we make decisions, what the relevant Community Standards are and if the user would like to accept or appeal the decision.

We also outline this process for users in our Transparency Centre, here:
https://transparency.fb.com/en-gb/enforcement/taking-action/taking-down-violating-content/#takedown-experience

**MODERATION**

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

Improvements that can be made to a service's content moderation are likely to vary depending on a range of factors, including the nature of the service, the extent to which it uses automated vs. human moderation, user numbers and demographics, and cost considerations for the provider. Meta is unable to comment on improvements that could be made to content moderation on other providers' services.

**12. What automated moderation systems do you have in place around illegal content?**

As explained in our responses above, our content moderation focuses on identifying violations of our Community Standards, which may overlap with various types of content that is illegal in the UK and legal content that is objectionable or harmful.

To enforce our Community Standards, we employ a combination of human review and technology. Every day, we remove millions of violating pieces of content and accounts on Facebook and Instagram. In most cases, this happens automatically, with technology such as artificial intelligence working behind the scenes to detect and remove violating content. We set out additional details about our automated moderation tools below.

Automated moderation tools

We use three primary forms of technology to detect Community Standards violations, which are developed and trained in different ways as set out below:

- First, we use privacy-protective matching technology (sometimes referred to as content hashing or content digital fingerprinting) to identify identical or near identical copies of URLs, text, images, audio and videos which we have previously identified as violating our Community Standards. This matching technology can work even if minor modifications have been made to the original content. When we match the content exactly or we determine it is near identical to previous violating content, we will typically remove the content.

The lists of known violating content that we use to power the matching technology are typically created after the same content has been repeatedly labelled as violating by our human reviewers. In addition, please see our response to question 19 below for details on how we use these technologies to help detect, remove, and report the sharing of images and videos that exploit children.

- Next, we employ rate limits (speed limits), which restrict the speed at which accounts can take actions on our platforms (e.g. making posts), to prevent misuse of bots.

  We set our rate limiting thresholds by observing how people use the platform and then setting conservative thresholds that allow us to address the worst bot behaviour while only infrequently affecting legitimate behaviour.

- Finally, we use artificial intelligence (in a narrow sense, i.e. machine learning and rules-based systems) in two ways. First, we use artificial intelligence to assess the likelihood of content violating our Community Standards. When confident enough that content violates our Community Standards, the artificial intelligence will typically remove the content. We also use artificial intelligence to select content for human review, based on severity, virality and likelihood of a violation of our Community Standards. As with the matching technology described above, artificial intelligence operates on URLs, text, images, audio and videos. However, unlike technologies that can only match violations they've seen before, artificial intelligence has the potential to identify certain violations it has never seen before.

The development of techniques used to train machine learning models is a fast moving area of study by industry and academia. Primarily, Meta uses two techniques to train its machine learning models.

The first technique is largely referred to as supervised learning. Meta's models for content moderation use variations on the same general technique for training these supervised models. Meta selects a statistically random sample of all content that users have viewed, which is the same method we use when we calculate our publicly reported prevalence of violations measurements, or a statistically random sample of all reports by our community. Human reviewers label the selected content as either benign or violating one or more of our Community Standards. As part of this process, the same content may be reviewed multiple times for quality control. We then combine these benign and violating examples as inputs into machine learning training algorithms. The output of these machine learning training algorithms is called a "model", often referred to as a "classifier". We can then use this classifier to determine if a post is likely to violate our Community Standards.

The second technique is referred to as self-supervised learning. In this training technique, the machine learning model removes a word from a sentence and then attempts to see if it can predict the missing word. This is a recently developed technique that Meta uses in more limited contexts, primarily to train language machine learning models.

Quality evaluation for automated moderation tools

Meta has ongoing quality evaluation processes to maintain and improve the accuracy of our automated content moderation tools (i.e. how accurate those tools are in detecting violations of our Community Standards).

Prior to fully launching any new rate limit, matching technology, or artificial intelligence tool, we use the technology to identify what it would delete, without actually deleting the content identified. We then use human reviewers to determine the tool's accuracy rate against real time content (rather than against historical content only, as is done when training the relevant technology). We find that these tools are often able to achieve higher levels of accuracy than human reviewers.

After launching rate limits, matching technologies, or artificial intelligence tools, we monitor the volumes of removals and objections by the user who posted the content, as well as the rate at which objections are granted. If any of the metrics we monitor appear to be showing abnormal signals, our engineering teams investigate. For example, if there are abnormal signals in the metrics used for an artificial intelligence tool, we will either send a sample of the artificial intelligence tool's recent results for human review to confirm the accuracy rate or deprecate the artificial intelligence tool.

In addition, many of our machine learning classifiers are automatically reassessed for accuracy after each human review, with the content labelling decisions taken by human reviewers being used to train and refine our technology. This helps to improve the quality of our artificial intelligence algorithms and the lists of known violating content used by our matching technology.

Use of such quality controls helps to ensure that automated removals adhere to our Community Standards, the development of which involves stakeholder consultations (including active engagement with NGOs, governments, individual activists, and academia) and extensive analysis of our internal signals, such as user research, large community surveys, and detailed analysis of what our community is reporting via platform reporting mechanisms.

We may also consider whether there are different algorithms that are better suited to address certain types of content or media manipulations and may use these algorithms in combination where needed. For example, in an effort to scale work done by Meta's fact-checkers related to Covid-19 misinformation, Meta deployed SimSearchNet, a convolutional neural net–based model built specifically to detect near-exact duplicates. Once independent fact-checkers have determined that an image contains misleading or false claims about coronavirus, SimSearchNet, as part of our end-to-end image indexing and matching system, is able to recognise near-duplicate matches so we can apply warning labels. Further information on these efforts is available here:

https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/

In addition, we have a Responsible AI team, which is a multidisciplinary team composed of ethicists, social and political scientists, policy experts, artificial intelligence researchers and engineers. We developed this team to help ensure that AI governance is based on foundational values of respect for human rights, democracy, and the rule of law. The team's overall goal is to develop guidelines, tools, and processes to help promote fairness and inclusion in AI at Meta, and make those resources widely available across the entire company so there is greater consistency in approaching questions of AI fairness. The team's efforts are organised around the key pillars of privacy and security, fairness and inclusion, robustness and safety, transparency and control, and accountability and governance.

The Data Transparency Advisory Group (DTAG), an independent body made up of international experts in measurement, statistics, criminology and governance, has also reviewed the question of whether we are accurately identifying content, behaviour and accounts that violate Facebook's Community Standards. They concluded that our process — combining automated and human review — is appropriate given the scale at which we operate and the amount of content people post. Moreover, the group found the way we audit the accuracy of our content review system was well designed, if executed as described. In their report, the group recognises the technical challenges that must be balanced in building an effective detection and enforcement system at scale. In particular, they note our current systems include mechanisms for people to report content to us and systems to detect harmful content even if someone hasn’t reported it. They recommended we bring more transparency to both processes and build additional ways for users to provide input into the policy development process itself.

Finally, Meta has established an independent Oversight Board, which can issue recommendations about the enforcement of our policies, including enforcement through automated means.

Developments in Meta’s use of artificial intelligence for content moderation

Addressing content that violates our Community Standards and Guidelines is one of the top priorities at Meta AI. Over the past five years, AI has become one of the most effective tools for reducing the prevalence of violating content, defined as the amount of violating content that people actually see on our platforms globally. AI systems have typically been single-purpose, each designed for a specific content type, language, and problem, like detecting misinformation or flagging hate speech violations, and they require varying amounts of training data and different infrastructure. Groups of bespoke systems result in high compute resources and maintenance complexity, which slows the process of updating systems to quickly address new, evolving challenges. But today, one of the biggest opportunities in our integrity work is to build not more bespoke AI systems but fewer, more powerful ones.

AI models that can combine signals across multiple systems help AI make new connections and improve content understanding. This also makes integrity systems more efficient by making better use of computer resources — which, crucially, allows us to respond more rapidly to new issues.

More broadly, generalisation is the path toward more intelligent AI systems that mimic the way humans learn. Rather than treating different tasks as completely separate, our brains can look at an object or piece of content and instantly make connections in ever-changing contexts. Teaching machines to do this well is one of the hardest and most important opportunities in AI.

As violating content continues to evolve and people look for new ways to evade our systems, we’ll continue our work to build more generalised AI systems that can adapt as needed to help keep people safe on our platforms.

For more information, see here:
https://ai.facebook.com/blog/the-shift-to-generalized-ai-to-better-identify-violating-content/

### 13. How do you use human moderators to identify and assess illegal content?

Please see our response to question 12, which outlines our use of both automated moderation tools and human reviewers as part of our content moderation system, and our response to question 15, which outlines how we deal with reports of illegal content. While AI systems are constantly improving, there will always be a need for human review, especially given the rapidly changing nature of harm online and the consequent evolving nature of Meta’s Community Standards and Community Guidelines.

In those instances where our technology does not flag relevant content or needs more input, Meta relies on human moderators to enforce our Community Standards and Community Guidelines. We have over 40,000 people working on safety and security, including 15,000 content reviewers around the world, so we can review reports across time zones. Our content review teams operate 24/7/365 and moderate in over 70 languages.

Our review teams review a blend of user reports and posts surfaced by our artificial intelligence tools. Our technology also supports the review teams by prioritising the most critical content to be reviewed, based on severity, virality and likelihood of a violation. Our review systems use technology to prioritise high-severity content with the potential for offline harm (e.g. posts related to terrorism and suicide) and viral content which is spreading quickly and has the potential to reach a large audience, in order to prevent as much harm as possible. This helps our human reviewers to focus on the most important borderline cases that require human review, which we also use to train our technology and strengthen our entire moderation system.

As potential content violations get routed to review teams, each reviewer is assigned a queue of posts to individually evaluate. Sometimes, this review means simply looking at a post to determine whether it goes against our policies, such as an image containing adult nudity, in instances when our technology didn't detect it first.

In other cases, context is key. For example, our technology might be unsure whether a post contains bullying, a policy area that requires extra context and nuance because it often reflects the nature of personal relationships. In this case, we'll send the post to review teams that have the right subject matter and language expertise for further review. If necessary, they can also escalate it to subject matter experts on the Global Operations or Content Policy teams.

When necessary, we also provide reviewers with additional information from the reported content. For example, words that are historically used as racial slurs might be used as hate speech by one person but can also be a form of self-empowerment when shared by another person, in a different context. In some cases, we may provide additional context about such words to reviewers to help them apply our policies and decide whether the post should be left up or taken down.

Meta's review teams consist of full-time employees who review content as part of a larger set of responsibilities, as well as content reviewers employed by our partners. They come from different backgrounds, reflect our diverse community and have an array of professional experiences – from veterans to legal specialists to enforcement experts in policy areas such as child safety, hate speech and counterterrorism.

We partner with companies that employ over 15,000 reviewers who help in doing the job of reducing harm. Our review teams are global and review content 24/7. As an essential branch of our content enforcement system, review teams must have language proficiency and cultural competency to do their job well.

In order to do their job, review teams undergo extensive training to ensure that they have a strong grasp on our policies, the rationale behind our policies and how to apply our policies accurately. Reviewers spend at least 80 hours in training with a live instructor. From there, they have hands-on practice using a facsimile of the review system, so they can apply what they've learnt in a simulated environment. After this hands-on learning, reviewers get a report highlighting the areas where they apply our policies consistently and accurately and areas where they need more practice. To ensure that they're up to speed on the latest information, reviewers receive regular coaching, refresher sessions and policy updates.

The tools used by reviewers in their daily work include the following:

- Standardised access to the Facebook Community Standards, Instagram Community Guidelines and other resources.

- Powerful search function within our Community Standards, Community Guidelines and training materials.
- Standardised review layout for different types of reported content.
- Customisable interface within the review tool.
- Highlighting tool for slurs and dangerous organisations based on the region where the content is reviewed.
- Tooltips that explain the definitions of certain words and how they should be used to inform decisions.

Review teams base their decisions on the detailed policies set out in the Facebook Community Standards and Instagram Community Guidelines. In theory, two reviewers reviewing the same posts would always make the same decision, but judgements can also vary if policies are not sufficiently prescriptive.

This is why Meta strives to make our policies as clear and comprehensive as possible. To help with this, a sample of reviewer decisions are audited on a regular basis to ensure that we're consistently applying our policies or identifying areas where improvements can be made. We've also implemented a tool, which is designed to facilitate feedback from reviewers on policy, tooling and other improvement ideas.

## ACTIONING CONTENT AND SANCTIONING USERS

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

Meta takes a three-part approach to content on Facebook and Instagram: remove, reduce and inform.

<u>Remove</u>

We remove content that goes against our policies. Meta notifies users so that they can understand why we removed the content and how to avoid posting violating content in the future.

Meta uses a strike system to count violations and hold users accountable for the content that they post. Depending on which policy the violating content goes against, the user’s previous history of violations and the number of strikes the user has, the user’s account may also be restricted or disabled.

Whether Meta applies a strike depends on the severity of the content, the context in which it was shared and when it was posted. If the content was posted to a Page or Community the user manages on Facebook, the strike may also count against that Page or Community. Further, if the user manages a Community, Meta may also count violations that the user approves as strikes against that Community. In some cases, a violation may be severe enough, such as posting child sexual exploitation content, that Meta will disable the user’s account, Page or Community on Facebook, or the user’s account on Instagram, after one occurrence.

To ensure that Meta’s strike system is fair and proportionate, Meta does not count strikes on violating content posted over 90 days ago for most violations or over four years ago for more severe violations. Meta also does not count strikes for certain policy violations, including when someone shares their own financial information, which Meta removes to prevent fraud, or cases where Meta has extra context about the nature of the violation.

If Meta removes multiple pieces of content at once, without providing the user sufficient time to view takedown notices and learn our policies, Meta may also count them as a single strike. All strikes on Facebook or Instagram expire after one year.

Reduce

If content on Facebook or Instagram doesn't violate the Facebook Community Standards or Instagram Community Guidelines, but might still be problematic or otherwise low-quality, Meta will reduce its distribution.

For example, on Instagram, we publish Recommendations Guidelines that govern what content and accounts we recommend to people. Through those guidelines, we work to avoid making recommendations that could be low-quality, objectionable, or sensitive, and we also avoid making recommendations that may be inappropriate for younger viewers. Specifically, the guidelines cover:

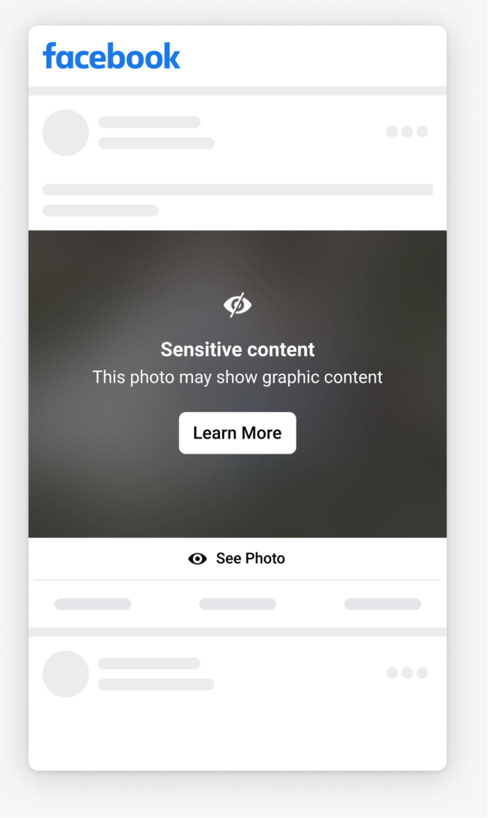
- content that impedes our ability to foster a safe community, such as content that may be sexually
- explicit or suggestive;
- sensitive or low-quality content about health or finance, such as content containing exaggerated health claims;
- content that users broadly tell us they dislike, such as contest promotions;
- content associated with low-quality publishing, such as news content lacking transparent information about authorship or the publisher's editorial staff;
- false or misleading content, such as vaccine misinformation that has been widely debunked by leading global health organisations; and
- accounts that, for example, have recently violated the Community Guidelines or have been banned from running ads on our platforms.

Our Recommendations Guidelines are designed to maintain a higher standard than our Community Guidelines, because recommended content and connections are from accounts that users haven't chosen to follow. We use technology to detect both content and accounts that don't meet these Recommendations Guidelines and to help us avoid recommending them, so that our users are less likely to encounter them.

Inform

One way Meta promotes a safe, authentic community is by informing people that content might be sensitive or misleading, even if it doesn't explicitly violate the Facebook Community Standards or Instagram Community Guidelines. In this instance, we'll include additional context about the content to help people decide what to read, trust or share.

For example, on Facebook, we include a warning screen over potentially sensitive content, such as violent or graphic imagery, posts that contain descriptions of bullying or harassment, if shared to raise awareness, some forms of nudity, and posts related to suicide or suicide attempts. An example of the warning screen is below.



On Instagram, we limit the visibility of certain posts that are flagged by people on Instagram for containing sensitive or graphic material. Photos and videos containing such content will appear with a warning screen to inform people about the content before they view it. This warning screen appears when viewing a post in feed or on someone's profile.

### 15. In what instances is illegal content removed from your service?

Generally, content on Facebook and Instagram that violates the Community Standards or Community Guidelines is taken down. In some cases, we may take other enforcement actions like labelling or limiting distribution or adding sensitivity screens.

Where content on Facebook or Instagram is reported as allegedly violating local law, we have a robust process for reviewing such reports.

When we receive a report, we first review it against the Facebook Community Standards or Instagram Community Guidelines. If we determine that the content goes against our policies, we remove it. If content does not go against our policies, in line with our commitments as a member of the Global Network Initiative and our Corporate Human Rights Policy, we conduct a careful legal review to confirm whether the report is valid, as well as human rights due diligence.

In cases where we believe that reports are not legally valid, are overly broad, or are inconsistent with international human rights standards, we may request clarification or take no action.

Where we do act against organic content on the basis of local law rather than our Community Standards, we restrict access to the content only in the jurisdiction where it is alleged to be unlawful and do not impose any other penalties or feature restrictions. We also notify the affected user.

We publish country specific information on content we restrict based on local law, which is available here: https://transparency.fb.com/data/content-restrictions/.

For example, in H2 2021, we restricted access in the United Kingdom to:

- two items, in response to reports from Her Majesty's Prison and Probation Service (HMPPS) and the National Offender Management Service (NOMS) that the content violated local laws prohibiting the use of mobile devices and sending information electronically from within a prison;
- 1,149 items in response to Consumer Policy reports submitted by various bodies, including the Advertising Standards Authority, the Medicines and Healthcare products Regulatory Agency, the Committee of Advertising Practice, the Gambling Commission, and the Financial Conduct Authority;
- 110 items in response to private reports of defamation;
- seven items for violations of local law; and
- one item in response to litigation.

While uncommon, we will occasionally receive legal demands that assert extraterritorial jurisdiction, and request that we restrict the availability of content globally.

While we respect the law in countries where we operate, we strongly oppose any extraterritorial legal demands and actively pursue all available options to appeal such orders. As an example, in H2 2021, we restricted access globally to 21 items in response to a court order from Brazil’s Supreme Court. We have actively pursued all options to appeal the orders.

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

We use a range of measures to address content that violates our standards, as set out in the responses to questions 12 and 13 above. These measures may result in the removal of violating content or in other actions that reduce its visibility and impact.

### 17. What other sanctions or disincentives do you employ against users who post illegal content?

Sanctions applied

As outlined in the answer to question 14 above, users who repeatedly violate the Facebook Community Standards or the Instagram Community Guidelines may lose access to some features for set periods of time, such as posting, commenting, using Facebook Live, or creating a Page or a Group.

Pages and Groups on Facebook that repeatedly violate our policies may be removed from recommendations and have their distribution reduced. Pages may also be restricted from certain monetisation features, and Groups may be required to have the admin approve posts.

For most violations, if the user continues to post content that goes against the Facebook Community Standards or Instagram Community Guidelines, despite repeated warnings and restrictions, Meta may disable the account. Meta may also remove Pages and Groups that repeatedly violate the Facebook Community Standards.

Preventing reaccess

As part of our Community Standards on Account Integrity, our recidivism policy prevents users from creating new accounts to evade our previous enforcement actions. Specifically, users cannot create or manage a new account, Page, or Group when we have previously removed their account, Page, or Group. This policy aims to prevent users from gaming our policies and their enforcement by disallowing creative reinterpretations of entities that we previously confirmed violated our policies.

We use a combination of human and automated review to enforce against recidivist accounts. We also take other steps to prevent reaccess in certain circumstances. For example, as part of our enforcement efforts against Dangerous Organisations and Individuals (DOIs), we undertake certain strategic network disruptions, to help make sure that these groups cannot find ways back on our platforms and proliferate on our platforms. As outlined in a recent Community Standards Report (https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/), between October 2019 and August 2020, we conducted 14 strategic network disruptions to remove 23 different organisations designated under our DOI Policy, including three U.S.-based white supremacist groups in October 2019.

Information provided to users

Whenever Meta removes content that goes against the Facebook Community Standards or Instagram Community Guidelines, the user may be notified so they can understand why the content was removed and avoid posting violating content in the future. Usually, this notice appears in the Feed when the user logs in to Facebook, or in their Feed on Instagram. Users will also be able to find this notice in their Support Inbox on Facebook or Support Requests on Instagram.

We'll also notify the user when a removal results in a restriction to their account, Page or Group, as a result of repeat violations of the Community Standards on Facebook, and Community Guidelines on Instagram.

Users can also access their history of violations, restrictions that their account might have and how long they'll last in the 'Account Status' feature on Facebook and Instagram. If a user manages a Page or Group, they can find information about violations and restrictions by looking at Page Quality or Group Quality.

Safeguards

As explained in our response to question 14 above, to ensure that subsequent restrictions are applied fairly, Meta won't count strikes on violating content posted over 90 days ago for most violations or over four years ago for more severe violations (such as Child Safety violations). Meta also won't count strikes for certain policy violations, such as when someone shares their own financial information, which we remove to prevent fraud. If we remove multiple pieces of content at once, we also count them as a single strike. All strikes on Facebook or Instagram expire after one year.

Users are also able to appeal the decisions we make regarding their content, accounts, Pages and Groups. To appeal a decision on Facebook or Instagram, people select the option to "Request Review" after we notify them that their content has been removed . When a review has been requested, Meta will review the post again and determine whether or not it follows our Facebook Community Standards or Instagram Community Guidelines.

If a user's Facebook or Instagram account has been disabled, they'll see a message saying that their account is disabled when they try to log in. In most instances, users are able to appeal the decision and request another review of the decision that has been made to disable their account.

## DESIGN AND OPERATION OF THE SERVICE, INCLUDING FUNCTIONALITIES AND ALGORITHMS

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

Meta has designed its services to be safe for all users, including in relation to illegal content, by providing a suite of in-app safety tools and privacy and security features available to all users service-wide. Our response to question 12 above outlines our use of automated moderation systems, which are designed to contribute to user safety by minimising the presence of illegal content and other violating content on our services. A number of additional examples are set out below.

Safety & Security

1. *Safeguards against child exploitation*. Meta maintains a zero tolerance policy when it comes to sharing sexual content involving teens on the Facebook and Instagram services. Meta's extensive efforts to combat child exploitation focus on preventing abuse, detecting and reporting content that violates its policies, and working with experts and authorities to keep children safe across the services. These efforts include:

    - *Photo-matching Technology*. Among the detection technologies used are photo-matching technologies that help detect, remove, and report the sharing of images and videos that exploit children. These photo-matching technologies create a unique digital signature of an image (known as a "hash") which is then compared against a database containing signatures (hashes) of previously identified illegal images to find copies of the same image. Meta uses these technologies across its public surfaces, as well as on unencrypted information available to it on its private-messaging services. Meta also runs these technologies on links from other internet sites shared on its services, and their associated content, to detect known child exploitation housed elsewhere on the internet. Not only does this help keep the Facebook and Instagram services safer, but it also helps keep the broader internet safer as all violating content is reported to NCMEC, who then works with appropriate law enforcement authorities around the world.

    - *Improving Detection Capabilities*. In addition to photo-matching technology, Meta uses artificial intelligence and machine learning to proactively detect child exploitative content when it is uploaded to the Facebook or Instagram services. Meta uses these technologies to more quickly identify this content and report it to relevant authorities (e.g., NCMEC), and also to find accounts that engage in potentially inappropriate interactions (e.g., "grooming") with children on the Facebook or Instagram services so that Meta can remove them and prevent additional harm.

    - *Child Safety Hackathon*. Meta continues to invest in its proactive detection technology to address child safety on its services. One such effort is by hosting a child safety-dedicated hackathon to bring together engineers and data scientists from across industry to develop technological solutions to help combat child sex trafficking. All code and prototypes developed for these hackathons are donated back to the Technology Coalition and Facebook's NGO partners to help them in their work to protect children. As part of the child safety hackathon, Meta recently committed to help fund the Internet Watch Foundation's initiative for young people to confidentially report self-generated sexual images of teens. Meta also recently committed to help fund a project led by Tech Matters that will develop new technology to support child helplines and make them more accessible to children in crises.

- *Focus on prevention*. Meta has developed targeted solutions, including new tools and policies to reduce the sharing of child exploitative content. For example, two new tools are aimed at preventing the searching and sharing of such content. The first is a pop-up that is shown to people who search for terms on the Facebook service associated with child exploitation. The pop-up offers ways to get help from offender diversion organisations and shares information about the consequences of viewing illegal content. The second is a safety alert that informs people who have shared viral meme child exploitative content about the harm it can cause and warns that it is against Meta's policies and there are legal consequences for sharing this material. Meta shares this safety alert in addition to removing the content, "banking it" (for the photo-matching technology) and reporting it to NCMEC. Accounts that promote this content will be removed.

- *Updated Reporting Tools*. After consultations with child safety experts and organisations, Meta made it easier to report content for violating its child exploitation policies. To do this, it added the option to choose "involves a child" under the "Nudity & Sexual Activity" category of reporting in more places on the Facebook and Instagram Services.

- *Project Protect*. Recognising that child exploitation is a problem across the internet and it is a collective responsibility to fight this abuse and protect children online, Meta has joined Google, Microsoft, and 15 other tech companies to announce the formation of Project Protect, a plan to combat online child sexual abuse. Project Protect focuses on five key areas: (1) tech innovations – to accelerate the development and usage of groundbreaking technology powered by a multi-million dollar innovation fund; (2) collective action – to convene tech companies, governments, and civil society to create a holistic approach to tackle this issue; (3) independent research – to fund research with the End Violence Against Children Partnership to advance a collective understanding of the patterns of child sexual exploitation and abuse online; (4) information and knowledge sharing – to continue to facilitate high-impact information and expertise; and (5) transparency and accountability – to increase accountability and consistency across industry through meaningful reporting, in conjunction with WePROTECT Global Alliance.

2. *Detecting and removing non-consensual intimate imagery (NCII) (or "revenge porn")*. Meta deploys technology to detect and remove NCII (using image processing and media match software). The algorithms look for indications of nudity or sexual activity in images and videos because they have been trained on previous violations of Meta policies. People over the age of 18 can also prevent the spread of their NCII by creating a case with StopNCII.org, which assigns a unique hash value (a numerical code) to their image, creating a secure digital fingerprint. Tech companies participating in StopNCII.org, like Meta, receive the hash and can use that hash to detect if someone has shared or is trying to share those images on their platforms. When creating a case with StopNCII.org, the original image never leaves the person's device. Only

hashes, not the images themselves, are shared with StopNCII.org and participating tech platforms. This feature prevents further circulation of that NCII content and keeps those images securely in the possession of the owner.

3. *Extensive support for potentially vulnerable users*. As part of Meta's efforts to keep people safe and address content that may be detrimental to the wellbeing of teens, Meta uses technical measures to proactively find and remove harmful suicide and self-harm content. This enables Meta to look for posts that likely break its rules around suicide and self-harm and make them less visible by down ranking them, and where Meta is confident that the content breaks its rules, remove that content from its platform. Meta uses this technology not just to find the content and make it less visible, but to send it to its human reviewers and get people help. Meta also provides anonymous reporting tools on both Facebook and Instagram for content such as self-injury posts. Meta may connect the account reported to organisations that offer help, as well as anonymous reporting for live videos to report at-risk behaviour during a live broadcast, so the person reported receives a message offering help, support, and resources. Global teams work 24 hours a day, 7 days a week, to review reports, provide this help, support, and resources. In addition, Meta provides support resources for users on topics such as suicide prevention, and information on who vulnerable users can reach out to in times of need. For instance, Meta provides a Safety Centre for suicide prevention, which contains resources developed by partners that are experts in this field and links to suicide prevention hotlines. These tools were developed in collaboration with mental health organisations such as Save.org, National Suicide Prevention Lifeline, Forefront and Crisis Text Line, as well as with input from people who have personal experience thinking about or attempting suicide.

4. *Messaging safeguards*. Meta deploys numerous safeguards for teen interactions on its direct messaging features:

    ○ *Restricting adults from direct messaging teens who don't follow them*. To protect teens from unwanted contact from adults, Meta prevents adults (18+) from sending messages to people under 18 who do not follow them. If an adult tries to message a teen who does not follow them, they will receive a notification that direct messaging is not an option.

    ○ *Safety notices*. If a teen and adult are already connected/following, Meta sends a safety notice prompt to encourage teens to be cautious in conversations. Safety notices in direct messages notify teens when a connected adult has been exhibiting potentially suspicious behaviour. For example, if an adult is sending a large amount of friend or message requests to people under 18, Meta will use this tool to alert the teen and give him/her the option to end the conversation, or block, report, or restrict the adult.

    ○ *Separate messaging folders*. On the Facebook service, unconnected user messages go into a separate mailbox where they can be deleted or reviewed without showing a read receipt to the sender. Messenger also gives users the option to ignore a conversation and automatically move it out of the inbox, without having to block the sender. On the Instagram service, users can control whether messages from different types of users (e.g., followers, others on Instagram, etc.) go into a "Message Request" folder, or choose not to receive them at all. All accounts on Instagram have the option to switch off direct messaging from people they don't follow. Users can also automatically hide direct message requests from people who don't follow them, or who only recently followed them.

5. *Minimising interactions between teens and adults*. Meta is making it more difficult for adults to find and follow teens on Instagram. This may include things like restricting adults from seeing teen accounts in "Suggested Users," preventing them from discovering teen content in Reels or Explore, and automatically hiding their comments on public posts by teens.

6. *Restricted features*. Teens generally have a more limited experience on the Facebook service when it comes to the features they have access to, who they share and connect with, and the content they see. For example, (1) Facebook Dating is available only to adults aged 18 years old and over who have a Facebook account in good standing; (2) only users who are aged 18 years old and over may offer or request mentorship in Facebook's mentorship groups; and (3) users must be aged 18 years old and over to buy and sell on Facebook Marketplace.

7. *Hiding likes and view counts*. Meta gives all users on Instagram and Facebook the option to hide likes and view counts on all posts in their feed and like counts on their own posts, so others can't see how many likes their posts got. See here for additional information: https://about.instagram.com/blog/announcements/giving-people-more-control

8. *Dedicated policies to address high-risk viral challenges*. To ensure the safety and security of its online communities and prevent real-world harm, Meta's content policy expressly prohibits "Coordinated Harm," explaining to users: "Do not post content that falls into the following categories … Depicting, promoting, advocating for or encouraging participation in a high risk viral challenge." Similarly, as stated in the Instagram Community Guidelines: "We're working to remove content that has the potential to contribute to real-world harm, including through our policies prohibiting coordination of harm . . . that contributes to the risk of imminent violence or physical harm." Meta has specialists within its Content Policy team that identify high-risk viral content (on and off the services). These specialists may identify a high-risk viral challenge by examining content reports on the services, through news sources or other information sources. This kind of attention ensures that Meta is prepared should a viral challenge begin on an online platform (or offline) and later appear on the Facebook and/or Instagram service(s). Meta deploys a number of techniques (including a combination of technology and human reviewers) to remove high-risk viral content that expressly violates its terms and policies and/or curtail its ability to spread across the services. For example, in addition to the reporting tools discussed above, Meta may curtail the use of "hashtags" that are often associated with a policy-violating, high-risk viral challenge.

9. *Nudges*. Instagram has started to deploy new "nudges" for teens. Currently, in certain countries, teens will see a notification that encourages them to switch to a different topic if they're repeatedly looking at the same type of content on Explore. This nudge is designed to encourage teens to discover something new and excludes certain topics that may be associated with appearance comparison.

10. *Anti-bullying tools and features*. With respect to an issue of particular relevance to teens, Meta has been seeking to lead in the fight against online bullying. Meta has put in place strong policies designed to provide heightened protection against bullying, provides anonymous reporting for bullying content, and has developed technology to detect and remove bullying content even before it is reported. These tools and features include:

- ○ *Bulk Blocking*. This tool helps users manage unwanted interactions by allowing them to easily delete comments in bulk, and block or restrict multiple accounts that post negative comments.

- ○ *Pinned Comments*. An easy way to give users the ability to amplify and encourage positive interactions by pinning a select number of comments to the top of their comments thread.

- ○ *Tag Controls*. Allows users to manage who can tag or mention them; located within their privacy settings under "Tags" and "Mentions," users can choose whether they want everyone, only people they follow, or no one to be able to tag or mention them in a comment, caption or Story.

- ○ *Blocking, unfriending, unfollowing*. When a user blocks, "unfriends" or "unfollows" another user, that other user is not notified. Meta also provides tools that allow users to easily limit their interactions with others, including limiting another user's ability to see their posts, and the posts they are tagged in, or limit who can see their past posts. On Instagram, Meta makes it harder for someone who a user has already blocked from contacting them again through a new account. With this feature, whenever a user decides to block someone on Instagram, they'll have the option to both block their account and preemptively block new accounts that person may create.

- ○ *Snoozing*. Users can use the "Snooze" feature on the Facebook service to stop seeing posts from certain people, Pages, or groups in their News Feed for 30 days; the person, Page, or group is not notified when they have been "snoozed" by another user.

- ○ *Mute*. Users are also able to "mute" interactions, to hide posts from certain accounts appearing on their Feed, without having to unfollow the account.

- ○ *Restrict*. On the Instagram service, users can enable the "Restrict" feature to "put some space" between themselves and another person's account, hiding the person's comments and messages on Direct (the private messaging feature on the Instagram service).

- ○ *Comment controls dashboard*. The Instagram service has a Comment Controls dashboard, easily located in users' privacy settings. Within the dashboard, users can choose to select "Allow Comments From" and allow comments from "Everyone," "People You Follow," "Your Followers," or "People You Follow and Your Followers." The "Block Comments From" option allows the user to block comments made by specified users. The dashboard also allows users to select "Filters." The "Hide Offensive Comments" filter uses machine learning to detect and automatically hide comments on the user's posts and live videos that may be offensive. Users can also select "Manual Filter," which allows them to create their own list of words or emojis they do not want to see in the comments section of their posts . They can also always use the in-app tools to remove individual comments from their posts, or turn off commenting altogether by selecting the "Turn Off Commenting" option.

- ○ *Limits*. To help protect people when they experience or anticipate a rush of abusive comments and direct messages, Meta introduced Limits on Instagram: a feature that is easy to turn on and automatically hides comments and direct message requests from people who are not followers or only recently started following a user. Limits allows a user to hear from his/her long-standing followers, while limiting contact from people who might only be coming to their account to target them with negativity.

- *Hidden Words*. The Instagram service has a Hidden Words dashboard, easily located in users' privacy settings. Hidden Words helps protect people from abuse in their comments and direct message requests; it allows a user to automatically filter offensive words, phrases, and emojis in comments and moves direct messages into a Hidden Folder, that a user never has to open if they don't want to. The user can select "Hide comments," "Hide message requests," or both. Users also have the option to create their own custom list of words, phrases and emojis they do not want to see in their comments and direct message requests.

- *Comment warnings to discourage harassment*. Meta shows a warning when someone tries to post a potentially offensive comment – reminding the person of the Community Guidelines and warning them that Meta may remove or hide their comment if they proceed. See here for additional information:

  https://about.instagram.com/blog/announcements/instagrams-commitment-to-lead-fight-against-online-bullying

- *Bullying Prevention Resources.* Meta has developed a dedicated Bullying Prevention Hub on Facebook and Anti-Bullying Centre on Instagram. Meta continues to explore and develop safeguards in this important space.

11. *"Sensitive Content" control*. Meta launched a Sensitive Content Control, which allows users to select whether to see less content that does not violate the Community Guidelines but that some users may not want to see or may find upsetting, offensive, or sensitive (e.g., content that may depict violence, such as people fighting; content that may be sexually suggestive; content that promotes certan regulatored products, cosmetic procedures, or sell health-related products and services) on different Instagram surfaces, like Explore, Search, Reels, Accounts You Might Follow, Hashtag Pages and In-Feed Recommendations. The Sensitive Content Control has three options: "More", "Standard", and "Less". "Standard" is the default state, and will prevent people from seeing some sensitive content and accounts. "More" enables people to see more sensitive content and accounts, whereas "Less" means they see less of this content than the default state. For people under the age of 18, the "More" option is unavailable, and existing teen users are sent a prompt encouraging them to select the "Less" option. New Instagram users under 16 years old will be defaulted into the "Less" state. These measures reduce the risk of teen users coming across potentially sensitive content and accounts.

12. *Time management tools*. The Instagram service offers tools to help teens (and families) better understand and manage how much time they are spending on the Instagram service and are a helpful tool to create healthy digital habits. These tools include: an activity dashboard; settings to activate reminders to take breaks and set daily limits; and a way to limit notifications.

    - The activity dashboard allows users to see how much time they spent on the service over the past day and week, as well as the average time they spent on the app.

    - Within the dashboard, they can set a reminder to help limit the time they spend on Instagram in a given day by creating an alert that will let them know when they have reached their time limit (e.g., every 15 minutes; 30 minutes; 45 minutes; 1 hour; or 2 hours).

- With "take a break," a user can set a cadence at which they would like to receive reminders to take a break if they've spent a certain amount of time at once on Instagram (e.g., every 10 minutes; 20 minutes; or 30 minutes). To ensure that teen users are aware of this feature, Meta served teen users notifications suggesting that they turn these reminders on.

- They can also silence push notifications for a select period of time.

13. *Parental Supervision*. Meta Platforms, Inc. first launched supervision tools on the Instagram service in the U.S. in March 2022, followed by the launch of additional Instagram supervision tools during a second launch 'wave' in June 2022, including to the UK, Ireland, France, and Germany in Europe, and a third launch 'wave' to all remaining locales where Instagram is available in September 2022. The current set of parental supervision tools allow parents and guardians whose teens opt in to or agree to use supervision to: (i) view how much time their teen spends on the Instagram service across devices in the last 7 days, (ii) set daily time limits, (iii) set scheduled breaks that limit a teen's use of Instagram during select days and hours, (iv) get notified when their teen shares that they have reported someone, and (v) view and receive updates on what accounts their teen follows and the accounts that follow their teen.

14. *Advertising policies*. Meta has strict advertising policies for advertising to all users, including teen users, which impose high standards on paid advertising. Among other things, the Advertising Policies and Branded Content Policies (applicable to the Facebook and Instagram services) strictly prohibit ads promoting the sale or use of certain types of products *for all users*, such as tobacco and related products, drugs and drug-related products, and adult content. Meta further age-restricts (i.e., 18+) ads for certain products or services, like alcohol, dating services, gambling, and weight loss products. These policies are actively enforced. For example, the Advertising Policies:

    - strictly prohibit ads promoting the sale or use of certain types of product, such as tobacco and related products, drugs and drug-related products, and adult content. See, e.g., sections under Prohibited Content entitled: "Tobacco and Related Products," "Drugs & Drug-Related Products," and "Adult Content." Simply put, this content cannot be advertised on the Instagram Service.

    - age-restrict ads for certain products or services, like alcohol, gambling, sexual and reproductive health products, dating services, and weight loss products – meaning that ads for such products and services must comply with all applicable local laws, including that such products/services may not be directed to individuals under 18.

    - provide that "*[a]ds targeted to minors must not promote products, services, or content that are inappropriate, illegal, or unsafe, or that exploit, mislead, or exert undue pressure on the age groups targeted*."

    - clearly explain that before an advertisement is delivered to any Instagram user, it is subject to an advertisement review which operates to detect violations of the Advertising Policies. This is to further reduce risk of delivery where, by accident or design, an advertiser creates advertisement that does not comply with applicable policies.

15. *Badge verification*. Instagram also uses verified badges to help people more easily find the public figures, celebrities and brands they want to follow. A verified badge is a check that appears next to an Instagram account's name in search and on the profile. It means Instagram has confirmed that an account is the authentic presence of the public figure, celebrity or global brand it represents, helping to reduce the risk of illegal activity involving the impersonation of such persons / brands.

16. *Encouraging use of our tools*. We are testing a new way to encourage teens to update their safety and privacy settings on Instagram. We'll show prompts asking teens to review their settings including: controlling who can reshare their content, who can message and contact them, what content they can see and how they can manage their time spent on Instagram. In addition, in the UK we recently ran a marketing campaign (https://www.youtube.com/watch?v=Pwpd7gBt5pM) – on our apps and elsewhere – to raise awareness and drive adoption of a number of tools we have launched to keep young people safe on our platforms, including our new parental supervision and time management tools on Instagram.

Privacy

1. *Limitations on ad targeting*. Meta applies teen-specific restrictions on ad targeting by default, known as the "slimmed down service." By default, users under 18 can only be targeted by age, gender, and location across Instagram and Facebook, and previously available targeting options – like interests or data from activity on third-party apps and websites – are no longer available to advertisers. Teens cannot opt-in to any such data use by advertisers for targeting purposes.

2. *Limitation on social interactions and ads*. For adults, Meta may include a user's social interactions alongside ads that a friend sees on the Facebook services. Social interactions include: page likes, app usage, and event responses. However, for users under 18 years old, their social interactions are not shown alongside any ads.

3. *Account audience defaults*. New teen accounts on the Facebook service are automatically defaulted to share their posts with "friends" only. This means only accounts that are a "friend" can view the content they share on the Facebook service, unless and until the teen changes their privacy settings. New teen accounts on the Instagram service are "private by default" meaning that during the new user flow, new teen accounts are presented with a screen containing two radial buttons ("private" or "public"), with the "private" option pre-selected by default and placed above the "public" option. For teens in the UK, Meta began defaulting new users on Instagram with a stated age under the age of 18 into a private account in July 2021. When this was rolled out, teens who were already on Instagram and who had a public account were shown a notification regarding private account controls that included the option to go to Settings to change their privacy settings. For private accounts, only approved followers can see the account's posts and stories. Teens who select "public" during the new user flow on the Instagram Service will receive a notification on the occasion of their first inbound follow or between 7 and 30 days after the user has completed account registration to remind them that they have the option to switch the audience setting for their personal account to "private" at any time. For existing teen users with a public account, Meta shows them an in-app notification to

remind them of their account privacy status, and explains how to change that setting. Teen users with a public audience setting will receive up to three of these notifications.

4. *Transparency*.

   - *Educational resources for teens*. Meta provides education for teens about its privacy features in the Youth Portal (https://www.facebook.com/safety/youth) (e.g., reviewing their timeline and tags, accessing their information, how ads work, and how to customise their privacy settings, including information on how to choose the audience for posts and how to take a privacy check-up). The Youth Portal also provides tailored and engaging information to help teens understand Meta's privacy policy. This youth-friendly privacy information helps to mitigate the risk that Teens may not understand how their personal data is processed. For Instagram users, Meta also offers a guide specifically aimed at teens, dedicated to staying safe online and creating a positive experience (https://www.instagram.com/instagram/guide/take-charge-create-a-positive-instagram-experience/17865134450117820/?igshid=1c2xjo6vbp91r). Meta also developed the Community Safety Centre (https://about.instagram.com/community/safety), which contains step-by-step instructions to guide them through using the privacy tools and features available on the Instagram service, links to additional resources, and programs to help them have a safe and positive experience.

   - *Educational resources for parents*. Meta also offers additional dedicated resources for parents, guardians, and other caregivers about the Facebook and Instagram services. These include a Parents Portal (https://www.facebook.com/safety/parents), Parent Centre (https://about.instagram.com/community/parents), and (as mentioned above) Parent's Guide (https://about.instagram.com/community/parents/guide), with information about the privacy and safety tools available to their teens on the Facebook and Instagram services, top questions from parents, and advice for talking to their kids about staying safe on Instagram.

   - *Family Centre*. Family Centre (https://familycenter.instagram.com/) is a place for parents and guardians (with their teens' permission) to oversee their teens' accounts within Meta technologies, set up and use supervision tools (discussed above), and access resources from leading experts. We have worked closely with experts, parents, guardians, and teens to develop the Family Centre. Meta's vision for Family Centre is to allow parents and guardians to help their teens manage experiences across Meta technologies, all from one central place.

   - *Education Hub*. The Family Centre also includes an Education Hub (https://familycenter.instagram.com/education/) where parents and guardians can access resources from experts and review helpful articles, videos, and tips on topics like how to talk to their teens about safe use of social media, which are available to access at any time. Parents can also watch video tutorials on how to use the supervision tools available to them. Meta worked closely with groups like Connect Safely and Net Family News to develop these resources, and will continue to update Family Centre's Education Hub with new information.

   - *Why am I seeing this ad?* For added transparency and control, the Facebook and Instagram services explain to users the basis for which they have been targeted for an ad in their feed. For teens, this would include the demographic categories chosen by the advertiser (given interest-based targeting is not available for teens).

**20. How do you support the safety and wellbeing of your users as regards illegal content?**

Please see question 19 above for a description of support that Meta provides for potentially vulnerable users. In addition, we work with expert organisations to build programmes that focus on helping young people with everything from bullying to providing parents with the tools to have conversations with the young people in their lives. For example: work with ParentZone and The Mix on "Digital Families" (https://www.themix.org.uk/digital-families), a toolkit and in person events that brings together families to discuss how they spend their time online; a partnership with The Diana Award to provide in school peer to peer training (https://diana-award.org.uk/anti-bullying/about/) on how to deal with bullying both online and offline; and working with Internet Matters on a guide for parents and young people on how to address the pressure to be perfect (https://www.internetmatters.org/wp-content/uploads/2020/02/AdviceForYoungPeople-A5-Booklet-1.pdf). Other examples of the support and resources we offer and our work with third parties are set out in our response to question 28 below.

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

Our responses to question 12 and 13 above outline a number of measures we take in relation to the design of various automated tools used in relation to our services.

**CHILD PROTECTION**

**24. Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?**

Meta's terms and policies set the minimum age to have a Facebook or Instagram account at 13 years old, as well as the Parents Portal, and several tools and technologies are used to enforce this minimum age policy and help to provide users 13+ with age appropriate experiences.

**25. If it is not possible for children to access your service, or a part of it, how do you ensure this?**

With specific service restrictions (such as Dating, Marketplace, Facebook Pay or the accounts memorialisation) that are only to be accessed by users aged 18 or over as well as the ad restrictions for advertisers willing to target their ads to users under 18, our services are available to children aged 13 and above. Please see our response to question 24 above regarding the steps we take in relation to controlling (and, where relevant, preventing) access by children.

**26. What information do you have about the age of your users?**

Please see our response to question 24.

## TRANSPARENCY

### 27. For purposes of transparency, what type of information is useful/not useful? Why?

Our Transparency Centre (https://transparency.fb.com/en-gb/) provides a hub for Facebook's and Instagram's integrity and transparency work, acting as a central destination for all updates on how we enforce our standards and how we respond to decisions, recommendations, and case updates from the Oversight Board. We publish regular reports to give our community visibility into how we enforce our policies, respond to data requests and protect intellectual property, while monitoring dynamics that limit access to Meta technologies. By updating the Centre regularly throughout the year (at least every half and quarterly in the instance of our flagship reports) we ensure that we provide consistently relevant and timely information to our stakeholders.

We have also established hubs for other aspects of our work, in order to ensure transparency and make it easier for our community to find relevant information. These include:

- The Privacy Centre (https://www.facebook.com/privacy/center), which sets out our Privacy Policy and provides guides to privacy-related topics and tools on Facebook and Instagram.
- The Safety Centres for Facebook (https://en-gb.facebook.com/safety/) and Instagram (https://about.instagram.com/safety), which bring together our overarching safety-related policies, tools and resources and provide hubs for our work in specific safety areas (such as women's safety, LGBTQ+ safety, suicide prevention, eating disorders and self-injury, and non-consensual intimate images).
- The Family Centre (http://meta.com/familycenter), discussed in our response to question 19 above, which is a place for parents and guardians (with their teens' permission) to oversee their teens' accounts within Meta technologies, set up and use supervision tools, and access resources from leading experts.
- The Parent's Guide to Instagram (https://about.instagram.com/community/parents), which gives parents resources to help their teens explore Instagram safely, including overviews of safety features, a glossary of Instagram features and settings, and conversation starters for discussions with teens about their experience on Instagram.
- The Facebook Parents Portal (https://www.facebook.com/safety/parents) and Youth Portal (https://www.facebook.com/safety/youth), which give parents and teens resources to understand Facebook's features, policies and tools, as well as tips and expert advice to help teen users have a safe and positive experience on Facebook.
- The Facebook Get Digital hub (https://www.facebook.com/fbgetdigital), which provides resources for teens, parents and educators to enhance digital citizenship and wellbeing and help young people become empowered in a digital world.

In addition, Meta believes openness and collaboration with the academic community based on voluntary arrangements will spur research and development, create new ways of detecting and preventing harmful content, and help keep people more safe. We are leading the way in collaborating with the academic community in various ways within the constraints of respecting the privacy of our individual users. We are open and transparent about our content moderation in three important ways: (1) the algorithms, (2) the individual results of those algorithms, and (3) aggregated results of all of our content moderation efforts.

Since our Facebook AI research (FAIR) lab (today: Meta AI) was founded in 2013, we have committed to an open science-based approach. Our research model revolves around publishing source code and methodologies, collaborating with other researchers across industry and academia, and creating open benchmarks and challenges. In addition, our affiliated researchers frequently publish the results of our source code and methodologies applied on our platforms. Some examples of our methodologies and technologies published in recent years include the following:

- XLM-R, Linformer, and RoBERTa: We have open-sourced our models and code so the research community can evaluate our natural language, especially multilingual, understanding machine learning models.
- Faiss, PDQ and TMK+PDQF: We have published our research and released the code for three of our algorithms used for finding identical and near identical copies of known photo and video content.
- CLARA: Confidence of Labels and Raters. We published our methodology and evaluation of boosting the accuracy of human reviewer labels.

Meta maintains Facebook Open Research and Transparency (FORT), a research and transparency initiative allowing controlled independent researcher access to datasets covering ads targeting, URL share activity, civic engagement and others. The URL share might potentially, to a certain degree, be used by researchers to identify potential gaps in our Community Standards and our enforcement, including automated enforcement.

As noted above, we publish our aggregated enforcement numbers across various areas of the Facebook Community Standards and the Instagram Community Guidelines in our quarterly CSER. For many violations, we publish measured viewership prevalence. Viewership prevalence is how many views of violating content we didn't prevent – either because we did not catch the violations early enough or did not identify them on Facebook and Instagram. We are always working to expand our prevalence coverage across violation types, regularly adding new ones each year as we refine and improve our methodologies.

But transparency is only helpful if the information we share is useful and accurate. In the context of the CSER, that means the metrics we report are based on sound methodology and accurately reflect what is happening on our platform. To this end, we worked with international experts in measurement, statistics, law, economics and governance to provide an independent, public assessment of whether the metrics we share in the CSER provide accurate and useful measures of Meta's content moderation challenges and our work to address them. They broadly agreed that we are looking at the right metrics and provided some recommendations for improvement. In August of 2020, we also committed to undertaking and releasing an independent, third-party assessment of our CSER and this year we delivered on that commitment by publishing EY's independent findings (available here: https://about.fb.com/news/2022/05/community-standards-enforcement-report-assessment-results/).
We also regularly update our methodologies where appropriate (an explanation of our processes for improving and checking our metrics is available here: https://transparency.fb.com/policies/improving/getting-better-at-measurement) and publish those changes (available here: https://transparency.fb.com/policies/improving/corrections-adjustments/).

This year, we also released our first annual Human Rights Report, covering 2020 and 2021, which sets out how we're addressing potential human rights concerns stemming from our products, policies or business practices. This report builds on the work we've done since 2018 of disclosing human rights impact assessments, as well as on a commitment we made in the Meta Human Rights Policy to report annually on our insights and actions from our human rights work. We've sought to ground the report, and our human rights work, in the United Nations' Guiding Principles on Business and Human Rights. We published the report here: https://about.fb.com/news/2022/07/first-annual-human-rights-report/.

## OTHER

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

In addition to the measures already described above, Meta takes a number of measures to enhance safety and well being as well as to provide support for those affected by harmful content – including unlawful content – distributed online. We set out examples below.

Meta works with a number of organisations and institutions which operate in the field of combating illegal content and promoting safety online. Meta also closely cooperates with law enforcement within applicable legal frameworks, who may use the dedicated Law Enforcement Online Request System (LEORS) for the submission, tracking and processing of requests. We also report the facts and circumstances of apparent or imminent instances of child exploitation appearing on our site from anywhere in the world to the NCMEC. NCMEC coordinates with law enforcement authorities from around the world.

Meta also offers comprehensive information and practical assistance to support affected people, which includes:

- Bullying Prevention Hub, a resource for teens, parents and educators seeking support and help for issues related to bullying and other conflicts.
- Stop Sextortion Safety Centre, a resource for anyone seeking support and information related to threats to reveal intimate images to get you to do something you don't want to do.
- Educational resources including safety tools with tips to stay safe in interactions with sharing, friending and reporting, securing your account, and protecting your information with specific sections for the safety of children, women, and the LGBTQ+ community. Through Get Digital!, our digital and well-being resource, we provide lesson plans, conversation starters and other resources to help young people become empowered in a digital world.
- An updated "Parent and Carer's Guide to Instagram" in the UK (https://about.fb.com/news/2022/02/launching-a-parents-and-carers-guide-to-instagram-in-ireland-for-safer-internet-day/), which provides parents and caregivers a deeper understanding of the safety and privacy features available to their teens on the Instagram service. Recognising that social media can sometimes feel confusing and potentially worrying for parents and caregivers, the guide also provides tips on how parents and caregivers can approach conversations about social media with their teens.

We also offer regular educational events in the form of trainings, workshops, webinars and in-person events for candidates ahead of elections, including safety tools and reporting flows as well as in relation to the fight against hate speech online.

Meta also formally launched the Courage Against Hate report (CAH Report), in partnership with the European Commission's Directorate-General for Justice & Consumers. Courage Against Hate (CAH) is an initiative brought together by Meta (at the time still "Facebook") for the purpose of sparking cross-sector, pan-European dialogue and action to combat hate speech and extremism. The CAH Report, published on 13 July 2021, brings together four research organisations (the Centre for the Analysis of the Radical Right, CARR; HOPE Not Hate; the Jena Institute for Democracy and Civil Society, Institut für Demokratie und Zivilgesellschaft, IDZ-Jena; and the Swedish Defence Research Agency/Uppsala University) and eight practitioner NGOs/companies (Iamhere International; Galop UK; Moonshot; the Media Diversity Institute/Textgain; the Institute for Strategic Dialogue, ISD; Zivilcourage &

AntiRassismus-Arbeit, ZARA) with the aim of helping to develop a mapping of both trends in hate speech and extremism as well as effective programs and initiatives countering hate in Europe. The report serves as a baseline for reference and counterspeech trainings by Meta's counterspeech NGO partners across Europe in 2022 and prompts a multi-disciplinary conversation around what policies, further analysis and programs are needed for the fight against hate, extremism and terrorism to be truly effective. This collection of articles unites European academic analysis with practitioners who are actively working on countering extremism within civil society, and demonstrates our continued commitment to tackling these issues.

We have also facilitated the launch of a website created by the UK NGO "The Revenge Porn Helpline (RPH)", which is available in a range of countries. RPH is an organisation that supports adult victims (i.e. individuals over the age of 18) of intimate image abuse and is a leading NGO in this space. The platform (StopNCII.org) is the first of its kind and RPH has designed it with the specific goal of empowering victims, by giving them a private and secure tool to proactively stop the proliferation of their non-consensual intimate image (NCII) online. It uses technology that hashes images and videos directly on the platform user's device, so that victims are not required to share the original content with either the NGO or the StopNCII platform. Once the hashes are shared with Facebook and Instagram, we use technology to detect identical or similar content as it is being uploaded to the platform and action that content accordingly.

Meta has also consulted a range of external groups in the development of children's wellbeing policies, including: (1) Family Online Safety Institute; (2) Center of Media and Child Health; (3) MediaSmarts; (4) the Yale Center of Emotional Intelligence; (5) the Fred Rogers Center; (6) the NCMEC; (7) Child Helpline International; (8) ECPAT; (9) INHOPE; (10) UNICEF; (11) Childnet International; (12) ConnectSafely; (13) FOSI; (14) Net Family News; (15) Centre for Social Research; (16) Telefono Azzurro; and (17) National Network to End Domestic Violence.

In relation to AI, we also invest in research and open-sourcing datasets and tools to help facilitate responsible use of AI, such as privacy-preserving machine learning, AI explainability, and fairness. For example, in 2021 we released our Casual Conversations data set, composed of over 45,000 videos designed to similarly help researchers evaluate computer vision and audio models for accuracy across a diverse set of ages, genders, apparent skin tones, and ambient lighting conditions.

We are also improving transparency by piloting simple, standardised documentation of our models and using interpretability software such as Captum. Although work in this area is still in its infancy, our hope is that ultimately we will be able to build an integrated transparency solution that can automatically feed information from internal documentation efforts into new transparency features and controls for the people using our products.

In addition to our technical research and product focused work, we are actively participating in efforts to establish clear AI principles and best practices, including collaborating with the OECD on the AI Observatory project to study and disseminate emerging best practices that are in line with its AI Principles.

Through our partnership with Open Loop, we are building innovative "policy prototyping" projects for testing new potential AI policy requirements with regulators and startups before they become law, to help ensure that they are both practical and impactful. We launched projects in Europe and the Asia Pacific region.

We are funding a global effort to solicit diverse academic research on AI ethics and governance topics, supporting the publication of academic papers in Asia, Africa, and Latin America and providing foundational support for an independent Institute for Ethics in Artificial Intelligence at the Technical University of Munich.