Aotearoa New Zealand Code of Practice for Online Safety and Harms

Meta's Annual Transparency Report

October 2023

Signatory:	**Meta Platforms, Inc.** Meta's 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. We 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. And we are building new ways to help people explore their interests and connect with the people they care about, including by building products and experiences for the metaverse.

If applicable: Relevant Products / Services:	**Facebook & Instagram** • 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 • Instagram brings people closer to the people and things they love. Creators can express themselves and push culture forward through a variety of ways, including photos, video, and connecting with and shopping from their favourite businesses.

Executive Summary:	Meta is proud to be a founding member and signatory to the Aotearoa New Zealand Code of Practice for Online Safety and Harms ("the Code"). This industry Code is a credible step in encouraging collaboration between the technology industry, civil society and government to combat online harms in a way that respects freedom of expression. In addition to this industry code, Meta is supportive of modernising laws and regulations for the digital age to reduce online harms, in ways that respect human rights. During the reporting period, some highlights of our work include: • Introducing of enhanced user controls, product features for parents and guardians and youth; • Launching new technology and partnerships to detect CSAM and support victims of sextortion; • Developing of Micro-learning modules - direct to classroom in New Zealand schools concerning online safety and media literacy; • Industry leading algorithmic transparency and user empowerment tools through the release of system cards, safety moderation support for smaller platforms through open sourced technology; • Disrupting co-ordinated inauthentic behaviour by foreign interference networks that impacted New Zealand, among other countries.

	• Combating misinformation, promoting transparency and ensuring election integrity on our platforms ahead of the New Zealand 2023 General Election. We look forward to continuing to work with New Zealand policymakers, civil society, academics and experts on steps to combat harmful content online in New Zealand over the next year.

Background	Meta's approach to online safety is consistent with our November 2022 baseline report (Baseline Report), which consists of five components: • **Policies** that provide clear rules on what is allowed and not allowed on our platforms • **Enforcement** processes, tools and technologies that helps us scale and accelerate policy enforcement efforts • **Tools, Products and Resources** that raise awareness of online safety issues, provide access to accurate and credible information, give more context on content in Feed, and provide people with more control over their online experience. • **Partnerships** that provide on-the-ground knowledge and expertise and enhance digital literacy education. • **Transparency** of our efforts for the public to scrutinise and hold us accountable. In the sections below, we summarise new measures introduced from July 2022 to July 2023 specific to the outcomes and measures, as outlined in section 4 of the Code. We mention both our localised initiatives and global partnerships and product updates where these have impacted New Zealand users. Meta has opted into every commitment under the Code. This present report should be read in conjunction with that Baseline Report given the latter outlines comprehensively our approach to combating online harms. Links are included throughout the report for further reference. It should also be read as a whole, given multiple initiatives in product, policies and partnerships will be relevant across different commitment areas and we have not repeated these to avoid duplication. In this background section, we outline some broader initiatives we launched to enhance safety, transparency and user empowerment, as well as initiatives that demonstrate our commitment to protecting users' right to free expression, safety, and privacy, all of which are guiding principles of the Code. **Meta Summit on Youth Safety** In December 2022, we held our first Meta Summit focused on Youth Safety and Well-Being to discuss the tools we have developed to support teens and families on our apps. The summit included safety advocates, mental health experts, educators, think tank researchers, policy writers and parents, and many others who helped inform the development of our tools to discuss challenges families face in the digital age and explore opportunities to better serve teens and families. In the past year, we have taken significant steps, including: • Developing parental controls that help parents and teens navigate their

time online together

- Using age verification technology to help teens have age-appropriate experiences
- Defaulting teens into more private settings
- Removing more content that violates our policies and making potentially sensitive content more difficult to find
- Helping to protect teens against unwanted interactions
- Offering tools for teens to spend more meaningful time online

More information on the summit and our youth safety tools can be found here.

Improving Facebook's Penalty System

Prompted by feedback from the Meta Oversight Board, we took steps to update Facebook's penalty system to make it more effective and fair.

Under the new system, we are focused on helping people understand why we have removed their content, which is shown to be more effective at preventing re-offending, rather than to quickly restrict their ability to post.

The changes are also fairer to those people who may have been disproportionately impacted by our old system, particularly when we made an incorrect moderation decision or missed context.

These changes, while based on our own analysis and feedback of the Oversight Board, are also responsive to feedback from our community — including our civil rights auditors — who noted that our old enforcement system needed more focus on proportionality. Independent experts who offered their guidance on this topic have routinely noted that our penalty system should have a better balance between punishing and encouraging remediation through education. .

More information on Facebook's updated penalty system can be found here.

Annual Human Rights Report

We're committed to respecting human rights in our business operations, product development, policies and programming. That commitment is embedded in our Corporate Human Rights Policy, which sets out how we apply the United Nations Guiding Principles on Business and Human Rights (UNGPs) to our apps and products, policies, programming, and overall approach to our business. As part of this, we release a public report annually on how we are addressing human rights concerns stemming from our products, policies or business practices — something very few other companies do. We also commission independent third-party experts to conduct human rights impact assessments on the role of our services in different countries.

Our first annual Human Rights Report was released in July 2022, covering 2020 and 2021. We sought to ground the report on the UN Guiding Principles on Business and Human Rights. The report detailed how we address human rights impacts, and included insights and actions from our human rights due diligence on products, countries and responses to emerging crises. It was built on the work we have been doing 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. The report covers:

- Updates to our policies as part of an ongoing effort to take human rights considerations into account in an increasingly dynamic world, including:
    - specifically referencing human rights principles;
    - clarifying our health misinformation policies;
    - enhancing our bullying and harassment policy to create stronger protections against gender-based harassment for everyone, including public figures;
    - prohibiting certain mass harassment or brigading;
    - expanding our policies that prohibit veiled and implicit threats, and more.
- How our Data Policy, our Law Enforcement Response Team and our due diligence assessments work together to protect people from unlawful or overbroad government surveillance.
- How we manage risks related to human trafficking and exploitation through in-product features that raise awareness, deter violating behaviour and offer support to victims.
- How our Community Standards and Community Guidelines address hate speech, how our advertising policies address non-discrimination, and how our dedicated Civil Rights and Human Rights teams worked across the company to help ensure responsible innovation and accessibility.
- Our work to increase teen safety on Instagram, our continuing work to fight child exploitation on WhatsApp, Facebook and Instagram and the significant investments we've made in teams and technologies to better protect free and fair elections, including dedicated teams focused on election integrity and products that bring people relevant and reliable voting information.

More details of our Human Rights Report can be found here.

Protecting Privacy and Integrity

We often talk about how we use data to maintain and improve the integrity of our technologies. To help explain what we mean by that, we published a white paper

in July 2023 that explains some of the ways we employ user data to create a positive experience across our technologies and the privacy considerations we take into account when utilising that data. The paper includes case studies that illustrate how Meta protects the privacy of our users, while also ensuring that they have a safe experience on our platforms. For example, some of the case studies cover:

- How we think about privacy when it comes to automated image processing and protecting the privacy of subjects of adult nudity
- How we retain data about people who have committed severe Community Standards violations to help prevent those people from starting new accounts
- How we verify people's identities while still protecting their privacy
- How we approach privacy, safety and security together in times of crisis when people may face threats and may lose access to our technologies

Aside from providing more transparency into our approach to privacy and data in relation to our safety, security and integrity, the paper aims to facilitate conversations s among privacy and content moderation experts, technology companies, and regulatory bodies on what we as a global community expect privacy to look like in the face of online safety, security, and integrity challenges.

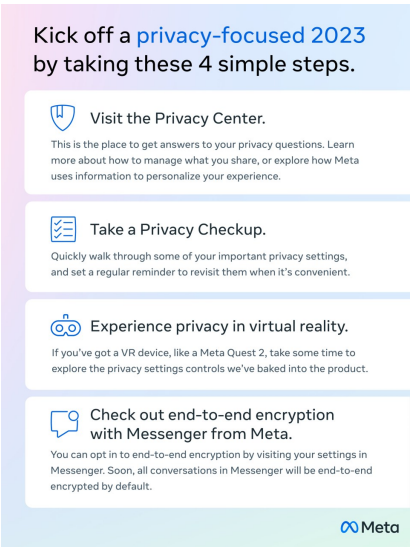
More information on the privacy and integrity white paper can be found here.

Building and Innovating With Privacy in Mind

In January 2023, we published the third iteration of our Privacy Progress Update and launched tools to help people better access the many privacy controls available in our apps. The following is a summary of our latest efforts:

- **Privacy Progress Update:** The third iteration of the report, which we publish each year to demonstrate technical and operational improvements we've made, includes details on privacy governance, consumer-facing privacy controls and innovative privacy solutions like privacy aware infrastructure.
- **Educational Tools:** We launched the Privacy Center alongside our easier to understand Privacy Policy in 2022 to give people a place where they could learn more about our approach to privacy across our apps and technologies. We also added two new modules to the Privacy Center — one focused on safety, where people can learn more about how they can protect themselves and their information; and one focused on privacy for teens, where people can learn about the special protections we offer for young people.
- **User Controls**: We notify people on Facebook to take a Privacy Checkup to review some of their important privacy settings. People can now set a

recurring reminder for Privacy Checkup to be automatically notified to use the control.



- **Protecting Teen Privacy:** In November 2022, we introduced updates on Facebook and Instagram that further protect teens online. Everyone who is under the age of 16 (or under 18 in certain countries) are now defaulted into more private settings when they join Facebook. In early 2023, we introduced new restrictions on the options advertisers have to reach teens, as well as the information we use to show ads to teens. Age and location are now the only information about a teen that we use to show them ads. We also added a new privacy page with more information for teens about the tools and privacy settings they can use across our technologies.

More information on these privacy initiatives can be found here.

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

Meta takes a comprehensive approach to child safety, including zero-tolerance policies prohibiting child sexual exploitation and abuse; technology to prevent, detect, remove and report policy violations; and victim resources and support. Our efforts to combat child exploitation focus on:

- Preventing exploitation and abuse of children with new tools and policies
- Detecting, removing and reporting exploitative activity that violates our policies
- Working with experts and authorities to keep children safe

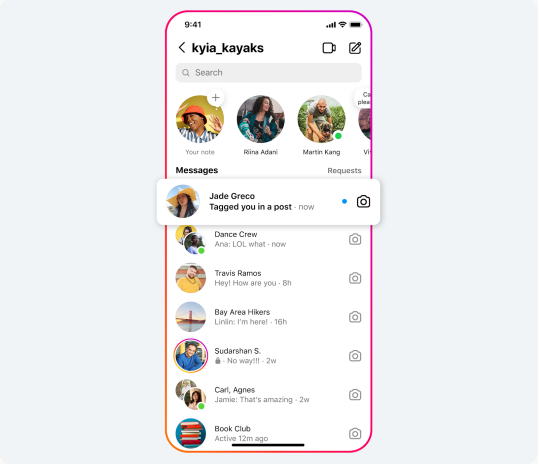
Further details can be found at facebook.com/safety/onlinechildprotection.

Combating exploitation of young people's intimate images

- In February 2023, we announced Instagram and Facebook as founding members of Take It Down — a new platform by the US National Center for Missing & Exploited Children (NCMEC) to find child exploitative content and help prevent young people's intimate images from being posted online in the future.
- Take It Down assigns a unique hash value — a numerical code — to their image or video privately and directly from their own device. Once they submit the hash to NCMEC, companies like ours can use those hashes to find any copies of the image, take them down and prevent the content from being posted on our apps in the future. Other companies can join to share hashes.

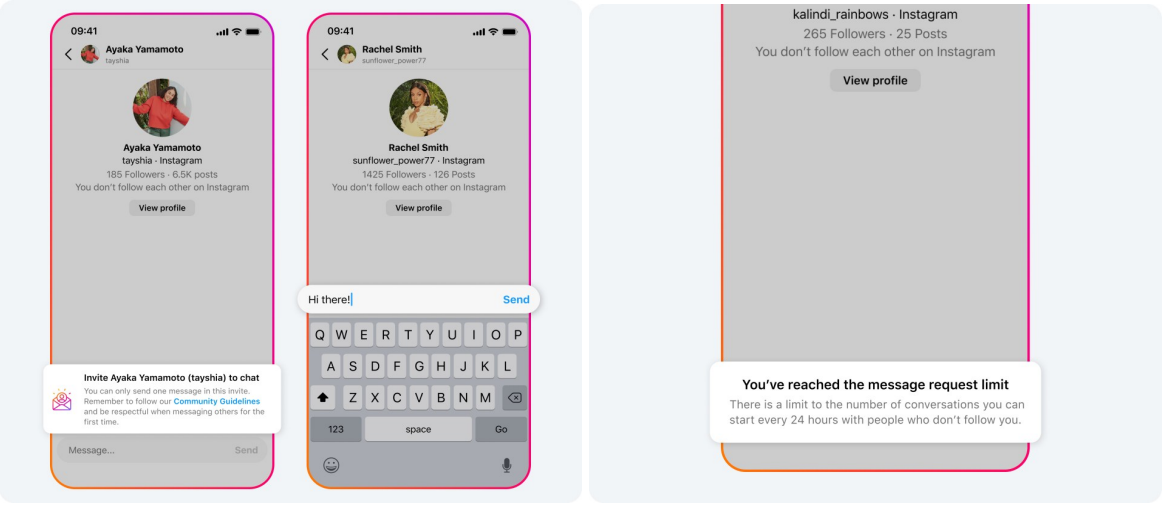
Increased Privacy for Teens on Instagram

- We work to prevent the posting of intimate images and sextortion as well as inappropriate interactions between young people and suspicious accounts attempting to take advantage of them. We now default teens into the most private settings on Facebook and Instagram, we work to restrict suspicious adult accounts from connecting with teens, and we educate teens about the dangers of engaging with adults they do not know online.
- On Instagram, adults are no longer able to see teen accounts when scrolling through the list of people who have liked a post or when looking at an account's Followers or Following list.
- If a suspicious adult account follows a teen account, we will send that teen a notification prompting them to review and remove the new follower. We also prompt teens to review and restrict their privacy settings. When someone comments on a teen's post, tags/mentions them in another post, or includes their content in Reels Remixes or Guides, the teen will receive a notification to review their privacy settings, and will have the option to stop people from interacting with them.

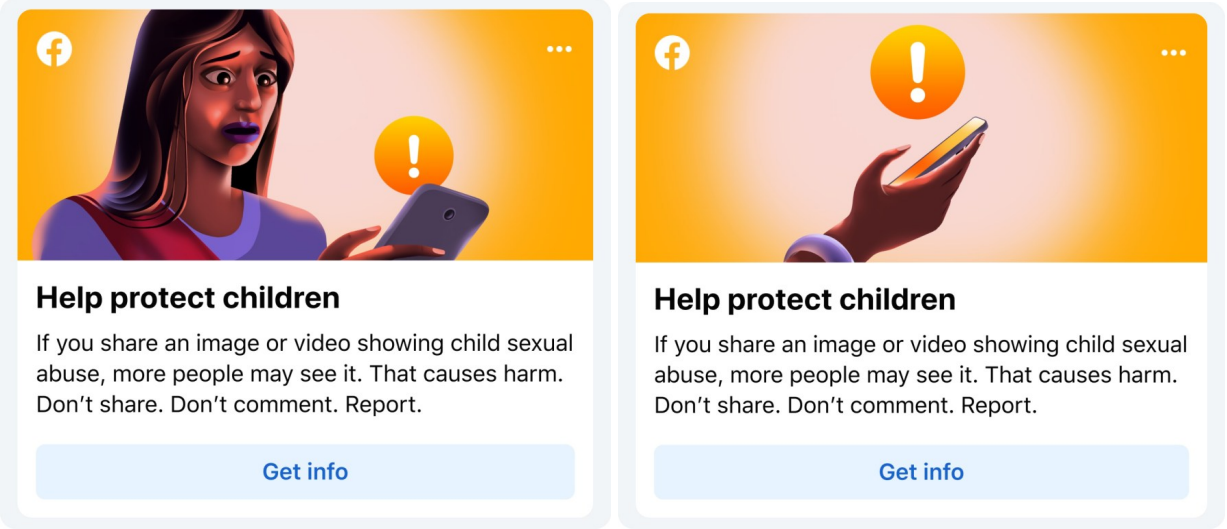


- We show Safety Notices when adults who have shown potentially suspicious behaviour message teens, and we restrict people over 19 years old from sending private messages to teens who don't follow them.

- Now, before being able to message someone who doesn't follow them, people must now send an invite to get their permission to connect. People can only send one invite at a time and can't send more until the recipient accepts the invitation to connect. We'll limit these message request invites to text only, so people can't send any photos, videos, or voice messages, or make calls, until the recipient has accepted the invite to chat. These changes mean people won't receive unwanted photos, videos, or other types of media from people they don't follow.



- In November, 2022 we launched a new partnership with Thorn and their NoFiltr brand to create educational materials that reduce the shame and stigma surrounding intimate images, and empower teens to seek help and take back control if they've shared them or are experiencing sextortion.



- Our research indicates that more than 75% of people that we reported to NCMEC for sharing child exploitative content shared the content out of outrage, poor humour, or disgust, and with no apparent intention of harm. Sharing this content violates our policies,

regardless of intent so we remove it. However, this also suggests more education is needed in this space so we ran a new education campaign encouraging people to stop before resharing those images online and report them instead.

Safeguarding Children New Zealand

- In September 2023, Safeguarding Children NZ held "Child Safeguarding Week" with a theme of preventing the sexual exploitation of children in New Zealand. We will further report on this initiative and our support of Safeguarding Children NZ in next years' report, but note here that Meta supported this week-long campaign (in some cases, ongoing, initiative) with financial, advertising and marketing expertise support expended over the reported period.

Global metrics for child endangerment content that we took action on globally in 2022 and the proactive rate of content we detected before people reported it.

Period	Child Nudity and Physical Abuse	Child Sexual Exploitation
Jan-Mar	Facebook: 2.1 million with proactive rate over 97% Instagram: 601,000 with proactive rate over 93%	Facebook: 16.5 million with proactive rate over 96% Instagram: 1.5 million with proactive rate over 92%
Apr-Jun	Facebook: 1.9 million with proactive rate over 97% Instagram: 481,000 with proactive rate over 93%	Facebook: 20.4 million with proactive rate over 99% Instagram: 1.2 million with proactive rate over 94%
Jul-Sep	Facebook: 2.3 million with proactive rate over 97% Instagram: 1,000,000 with proactive rate over 96%	Facebook: 30.1 million with proactive rate over 99% Instagram: 1.3 million with proactive rate over 96%
Oct-Dec	Facebook: 2.5 million with proactive rate over 98% Instagram: 621,000 with proactive rate over 97%	Facebook: 25.2 million with proactive rate over 99% Instagram: 9.7 million with proactive rate over 99%

For New Zealand, from January to December 2022:

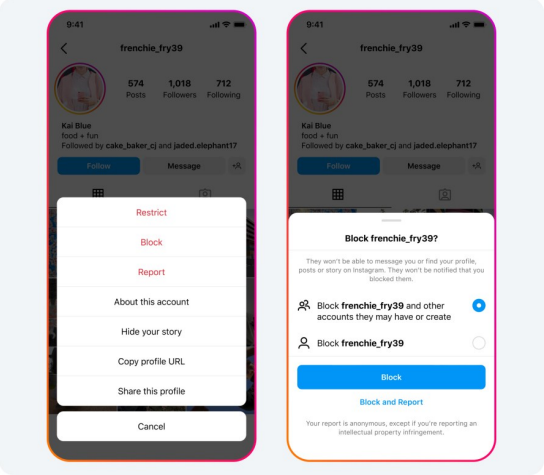
- **We took action on over 190,000 pieces of content on Facebook in New Zealand for violating our Child Sexual Exploitation policies**. Over 99% of this content was detected proactively before people reported it to us.
- **We took action on over 8,000 pieces of content on Instagram in New Zealand for violating our Child Sexual Exploitation policies**. Over 92% of this content was detected proactively before people reported it to us.

Note: In our Baseline report we had included aggregated metrics for Child Nudity and Sexual Exploitation, but due to the changes in our approach to the metrics for child endangerment content, we now separately report metrics for Child Nudity and Physical Abuse and Child Sexual Exploitation.

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

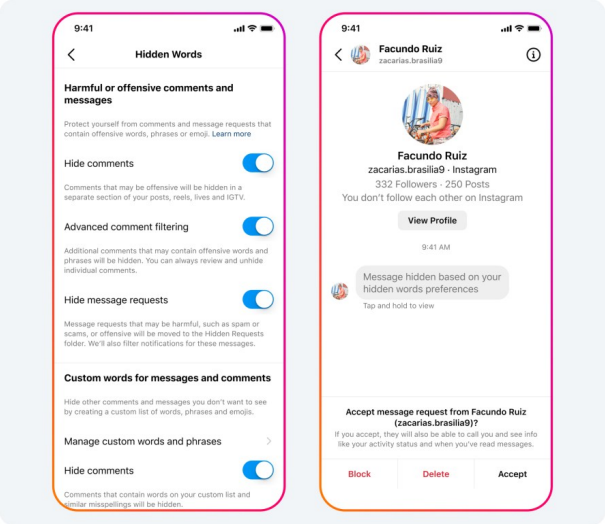
Updates to blocking newly created accounts:

- In October, 2022 we updated our tools on how we protect people from abuse on Instagram. When a user blocks someone, they now also have the option to block other accounts they may have or create, making it more difficult for them to contact the user.



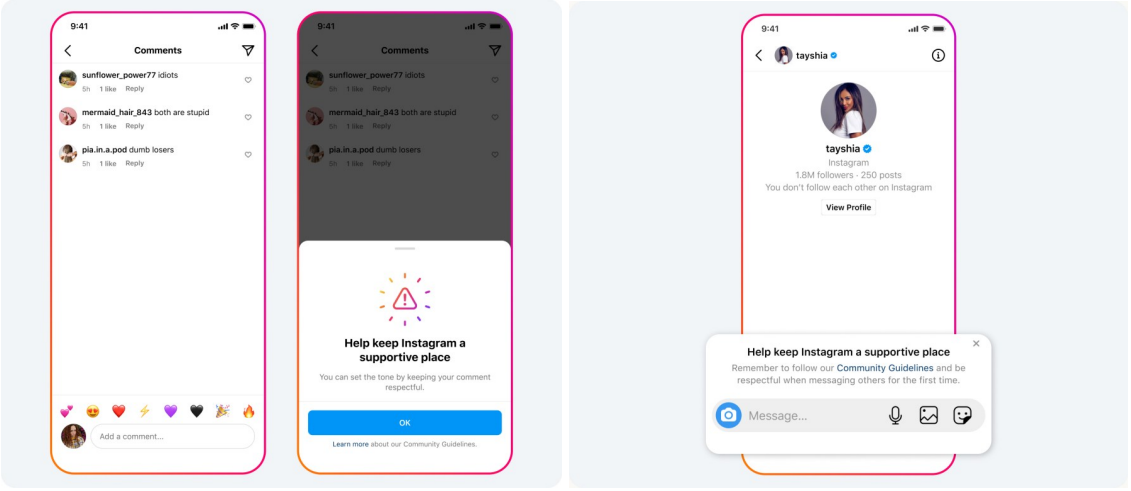
Updates to Hidden Words

- In 2021 we launched Hidden Words and can now see that one in five people with more than 10,000 followers have turned on the feature, giving them a powerful tool to automatically filter harmful content from their comments and message requests (they see 40% fewer comments that might be offensive).
- In 2022, we extended this to more creators to benefit from this protection, so we're now automatically turning on Hidden Words for Creator accounts. Everyone will continue to be able to turn these settings on or off at any time and build a custom list with additional words, phrases and emojis they may want to hide.
- We've also also expanded the Hidden Words feature to offer more protections, including:
    - to cover Story replies, so offensive replies from people you don't follow will be sent to your Hidden Requests folder and you never have to see them.
    - Improving our filtering to spot and hide more intentional misspellings of offensive terms, for instance, if someone uses a "1" instead of an "i."
    - Adding new terms to filter message requests that might contain scams or spam.



Nudging People to Be Respectful in Comments and DMs

- Building on new nudging prompts in 2021, in late 2022 a new notification began encouraging people to pause and consider how they want to respond before replying to a comment that our systems detects as potentially offensive.



- We now also remind people to be respectful in DMs when sending a message request to a creator. This nudge helps people remember that there's a real person on the other side of their DM request, and encourages more respectful outreach to people they may not know.

NetSafe New Zealand Partnership

- In New Zealand our Trusted Partner (a formal designation for external organisations we onboard to dedicated reporting channels) is Netsafe — New Zealand's independent, not-for-profit online safety organisation. We supported Netsafe initiatives in a number of ways over the reporting period, including:

- Holding research and youth engagement sessions on safety in the Metaverse, with practical demonstrations in focus groups in Dunedin and Auckland.
- Providing support to NetSafe's Stop Sextortion campaign, which included a call to action to report to both NetSafe and use the StopNCII tool (an international tool developed by Meta and outlined in the Baseline report).
- Supporting NetSafe as a founding member alongside Consumer New Zealand, IDCare New Zealand and the Global Anti-Scam Alliance, NetSafe's anti-scam activity including rolling-out the CheckNetSafe tool and opening scams reporting channels to NetSafe.
- Funded, developed and deployed a series of online modules - Micro Learns - to help educate young people, their families and educators on online safety and media literacy. The modules included:
  - Media literacy on social platforms
  - Staying Safe on Instagram - for young people
  - Owning your info on Facebook - for young people
  - Safety on Facebook and Instagram - for parents
  - Metaverse - staying safe in new digital spaces
  - Metaverse - an introduction for families
- NetSafe continues to be one of only 12 entities globally that sits on the independent Meta Safety Advisory Group.
- Over this Code's reporting period, we received over 1500 reports from NetSafe, flagging nearly 5,000 specific items. Of these, the top three violation types were (1) Bullying and Harassment (2) Adult Sexual Exploitation (3) Child Sexual Exploitation. We also received reports in New Zealand from the Department of Internal Affairs, New Zealand Police.

YouthLine Partnership

- We continue to provide YouthLine with free advertising support. In particular in 2023, we supported YouthLine to deploy resources and support awareness of support services to areas impacted in following Cyclone Gabrielle given the mental health impacts experienced in the Hawkes Bay / East Coast region following the devastation that occurred.
- *Note: In the 2022 Baseline report we reported work that we had undertaken with Youthline, expected to be delivered in 2023. Unfortunately due to unforeseen delays - which includes research into post-vention and a postvention toolkit (following incidences SSI, in an effort to combat a phenomenon knowns as 'clustering') - this initiative will now be delivered in late 2023 and will be outlined in detail in next year's report.*

Global metrics on bullying and harassment content that we took action on globally in 2022 and the proactive rate of content detected before people reported it.

Period 2022	Facebook	Instagram
Jan-Mar	9.5 million with proactive rate over 67%	7 million with proactive rate over 83%
Apr-Jun	8.2 million with proactive rate over 76%	6.1 million with proactive rate over 87%
Jul-Sep	6.6 million with proactive rate over 67%	6.1 million with proactive rate over 84%
Oct-Dec	6.4 million with proactive rate over 61%	5 million with proactive rate over 85%

For New Zealand, from January to December 2022:

- **We took action on over 49,000 pieces of content on Facebook in New Zealand for violating our Bullying & Harassment policy**. Over 62% of this content was detected proactively before people reported it to us.
- **We took action on over 89,000 thousand pieces of content on Instagram in New Zealand for violating our Bullying & Harassment policy**. Over 89% of this content was detected proactively before people reported it to us.

Outcome 3: Provide safeguards to reduce the risk of harm arising from online hate speech

- We don't allow hate speech on our platforms. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence. We define hate speech as attacks on individuals or groups with a list of protected characteristics, while continuing to allow people to criticise concepts, ideas and movements. Learn more about Meta's approach to hate speech here.
- Aside from the many partnerships we have around the world to combat hate, we have continued our support in New Zealand with the Sakinah Community Trust and their efforts to promote community cohesion. This includes supporting Unity Week. The Trust is a women-led organisation that focuses on the development of long-term community response and engagement following the 15 March 2019 attacks. Unity Week has been held annually since March 2022, and includes a range of free community events.
- In 2023, we supported Auckland Pride with free advertising to support their mission to raise awareness of empowering, celebrating, and serving Takatāpui & Rainbow Communities through events, creativity, and advocacy.

Global metrics on the pieces of hate speech content that we took action on globally in 2022 and the proactive rate of content detected before people reported it:

Period	Facebook	Instagram
Jan-Mar	15.1 million with proactive rate over 95%	3.4 million with proactive rate over 89%
Apr-Jun	13.5 million with proactive rate over 95%	3.8 million with proactive rate over 91%
Jul-Sep	10.6 million with proactive rate over 90%	4.3 million with proactive rate over 93%
Oct-Dec	11 million with proactive rate over 81.9%	4.7 million with proactive rate over 93%

For New Zealand, in 2022:

- **We took action on over 60,000 pieces of content on Facebook in New Zealand for violating our Hate Speech policy.** Over 92% of this content was detected proactively before people reported it to us.
- **We took action on over 43,000 pieces of content on Instagram in New Zealand for violating our Hate Speech policy.** Over 95% of this content was detected proactively before people reported it to us.

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

Freedom of expression is a foundational human right and enables many other rights. But we know that technologies for free expression, information and opinion can also be abused to spread hate and misinformation that serve as tools for the incitement of violence. This requires developing both short-term solutions that we can implement when crises arise and having a long-term strategy to keep people safe on our platforms.

Hasher-Matcher-Actioner (HMA)

- We are an industry leader in developing AI technology to remove hateful content at scale. As part of our wider commitment to protecting people from harmful content, we recently made available a free open source software tool we developed that will help platforms identify copies of images or videos and take action against them en masse. We hope the tool — called Hasher-Matcher-Actioner (HMA) — will be adopted by a range of companies to help them stop the spread of terrorist content on their platforms, and will be especially useful for smaller companies who don't have the same resources as bigger ones. This builds on our previous open source image and video matching software, and can be used for any type of violating content. Supporting smaller platforms is a key concern of the Christchurch Call to Action, and this tool is a significant contribution in this space.

Expanding our Dangerous Individuals and Organisations policy to address violence inducing social movements.

- We have expanded our Dangerous Individuals and Organisations policy to address organisations and movements that have demonstrated significant risks to public safety but do not meet the rigorous criteria to be designated as a dangerous organisation and banned from having any presence on our platform. While we will allow people to post content that supports these movements and groups, so long as they do not otherwise violate our content policies, we will restrict their ability to organise on our platform.

- Under this policy expansion, we have imposed restrictions to limit the spread of content from Facebook Pages, Groups and Instagram accounts. We also remove Pages, Groups and Instagram accounts where we identify discussions of potential violence, including when they use veiled language and symbols particular to the movement to do so. We take the following actions on Facebook and Instagram accounts, pages and groups associated with these movements, including:

    - Removal when they discuss potential violence.
    - Limiting recommendations making them ineligible to be recommended to people.
    - Reduce ranking in Feed.
    - Reduce in Search meaning they will not be suggested through our Search Typeahead function and will be ranked lower in Search results.
    - Reviewing Related Hashtags on Instagram: We have temporarily removed the Related Hashtags feature on Instagram, which allows people to find hashtags similar to those they are interacting with.
    - Prohibit Use of Ads, Commerce Surfaces and Monetization Tools:
    - Prohibit Fundraising:

- As of August 15, 2022, we have identified over 1,151 militarised social movements to date and in total, removed about 4,200 Pages, 20,800 groups, 200 events, 59,800 Facebook profiles and 8,900 Instagram accounts. We've also removed about 4,200 Pages, 12,000 groups, 840 events, 67,200 Facebook profiles and 38,800 Instagram accounts for violating our policy against QAnon.

Global Internet Forum for Countering Terrorism

- In January 2023, Meta assumed the chair of the Global Internet Forum to Counter Terrorism (GIFCT)'s Operating Board. GIFCT is an NGO that brings together technology companies to tackle terrorist content online through research, technical collaboration and knowledge sharing.
- Meta is a founding member of GIFCT, which was established in 2017 and was operationalised and evolved into a nonprofit organisation following the 2019 Christchurch Call, a movement we continue to support.
- In May 2023, we hosted a GIFCT and Meta Summit in Singapore focusing on combating online extremism. Our intention was to bring this conversation to an Asia-Pacific community and share insights and challenges amongst NGOs, governments and technology companies. We were eager to see New Zealand represented strongly at this session and

facilitated Sakinah Trust, NetSafe and Christchurch Call leadership from the New Zealand government to join us with all groups taking part in panels and leading conversations.

Global metrics on pieces of content that incite violence that we took action on globally in 2022 and the proactive rate of content we detected before people reported it.

Period	Facebook	Instagram
Jan-Mar	21.7 million with proactive rate over 98%	2.7 million with proactive rate over 95%
Apr-Jun	19.3 million with proactive rate over 98%	3.7 million with proactive rate over 97%
Jul-Sep	14.4 million with proactive rate over 94%	4.5 million with proactive rate over 97%
Oct-Dec	13.1 million with proactive rate over 87%	5.3 million with proactive rate over 97%

For New Zealand, in 2022:

- **We took action on over 147,000 of content on Facebook in New Zealand for violating our Violence & Incitement policy.** Over 95% of this content was detected proactively before people reported it to us.
- **We took action over 51,000 pieces of content on Instagram in New Zealand for violating our Violence & Incitement policy.** Over 98% of this content was detected proactively before people reported it to us.

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

- We know that people have different sensitivities with regard to graphic and violent imagery, which is why we have a multi-prong policy and enforcement actions to address different levels of sensitivities and situations.
- To protect users from disturbing imagery, in addition to other initiatives mentioned above, we remove content that is particularly violent or graphic, such as videos depicting dismemberment, visible innards or charred bodies.
- We also remove content that contains sadistic remarks towards imagery depicting the suffering of humans and animals. In the context of discussions about important issues such as human rights abuses, armed conflicts or acts of terrorism, we allow graphic content (with some limitations) to help people to condemn and raise awareness about these situations. There are also categories of content that we may allow on our platform for public interest, newsworthiness or free expression value, that may be disturbing or sensitive for some users.

Global metrics for pieces of violent and graphic content that we took action on globally in 2022 and the proactive rate of content detected before people reported it.

Period	Facebook	Instagram

Jan-Mar	26.1 million with proactive rate over 99%	6.1 million with proactive rate over 99%
Apr-Jun	45.9 million with proactive rate over 99%	10.1 million with proactive rate over 99%
Jul-Sep	23.2 million with proactive rate over 99%	6.9 million with proactive rate over 99%
Oct-Dec	15.5 million with proactive rate over 98%	6.1 million with proactive rate over 98%

For New Zealand, in 2022:

- **We took action on over 43,000 thousand pieces of content on Facebook in New Zealand for violating our Violent and Graphic Content policy.** 98% of this content was detected proactively before people reported it to us.
- **We took action on over 14,000 pieces of content on Instagram in New Zealand for violating our Violent and Graphic Content policy.** 97% of this content was detected proactively before people reported it to us.

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

- Our approach to misinformation is guided by the principle that we should provide people with accurate and informative content, while balancing free expression. Our users want to see high quality content on our platform, which is why our strategy to combat misinformation has three parts: remove, reduce, and inform (as noted above in our Baseline Report in 2022).
- Misinformation is a complex social phenomenon, which involves a range of offline and online behaviours, and goes beyond any single platform. Unlike the other types of harmful content addressed by this Code — there is no clear way to articulate what should be prohibited. There is an inherently fraught definitional challenge - governments, political entities, policymakers, civil society, academics, journalists and people in general do not agree on what misinformation is. What one person considers to be false or misinformation, may simply be another's opinion.
- Moreover, there is an important difference between misinformation shared unintentionally and misinformation shared intentionally to deceive - commonly referred to as "disinformation" (as described in the next section). Defining what constitutes misinformation is very challenging, but adding to the challenge is determining who decides if something is untruthful — who or what is the source of truth — which often comes with differing views.
- For the purpose of this report, the terms "misinformation" and "disinformation" are defined as:
    - Misinformation refers to *content* that is false or misleading;

- Disinformation refers to coordinated efforts to manipulate public debate for a strategic goal, with the intention to deceive, and involve *behaviour* that is inauthentic

COVID-19 Misinformation Policies

- In July 2022, Meta Asked the Oversight Board to Advise on COVID-19 Misinformation Policies. Specifically, Meta sought an advisory opinion from the Oversight Board to guide whether our COVID-19 misinformation policy is still appropriate given the evolving nature of the pandemic. Under this policy, we began removing false claims about masks, social distancing, vaccines and more. Given the evolution of the COVID-19 situation, we sought the Oversight Board's opinion on whether we should change the way we address this type of misinformation through other means, like labelling or demoting it.

- Misinformation related to COVID-19 has presented unique risks to public health and safety over the last two years and more. To keep our users safe while still allowing them to discuss and express themselves on this important topic, we broadened our harmful misinformation policy in the early days of the outbreak in January 2020. Before this, Meta only removed misinformation when local partners with relevant expertise told us a particular piece of content (like a specific post on Facebook) could contribute to a risk of imminent physical harm. The change meant that, for the first time, the policy would provide for removal of entire categories of false claims on a worldwide scale.

- As a result, Meta has removed COVID-19 misinformation on an unprecedented scale. Globally, more than 25 million pieces of content have been removed since the start of the pandemic. Under this policy, Meta began removing false claims about masking, social distancing and the transmissibility of the virus. In late 2020, when the first vaccine became available, we also began removing further false claims, such certain claims that the vaccine was harmful or ineffective. Meta's policy currently provides for removal of 80 distinct false claims about COVID-19 and vaccines.

- Meta remains committed to combating COVID-19 misinformation and providing people with reliable information. As the pandemic has evolved, we have adapted those policies introduced in the early days of an extraordinary global crisis.

- The Oversight Board was established to exercise independent judgement, operating as an expert-led check and balance for Meta, with the ability to make binding decisions on specific content cases and to offer non-binding advisory opinions on its policies.

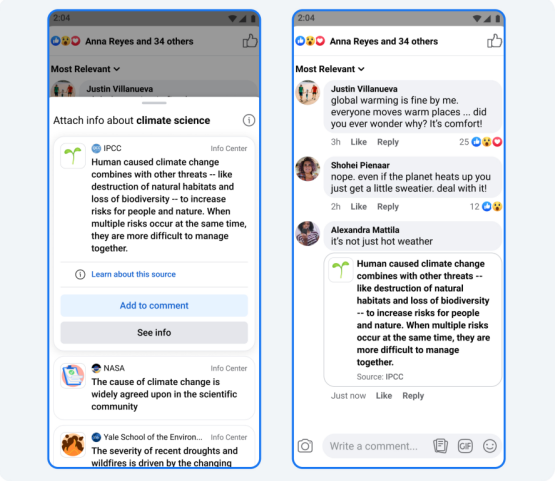
- In July 2023, we released our response to the recommendations the Oversight Board made in their Covid-19 misinformation Policy Advisory Opinion. We now take a more tailored approach to our Covid-19 misinformation rules consistent with the Board's guidance and our existing policies. In countries that have a Covid-19 public health emergency declaration, we will continue to remove content for violating our Covid-19 misinformation policies given the risk of imminent physical harm. We continue consulting with health experts to understand which claims and categories of misinformation could continue to pose this risk. Our Covid-19 misinformation rules are no longer in effect globally as the global public health emergency declaration that triggered those rules has been lifted. This includes New Zealand

where the Government lifted all remaining COVID-19 restrictions in August 2023, having moved out of a state of emergency system in September 2022.

- A key part of our approach to combat misinformation is providing tools and products that will contribute to a more resilient digital society, where people are able to critically evaluate information, make informed decisions about the content they see, and self-correct.

Example - Combating Climate Change

- We've added a Climate InfoFinder and Climate Science Literacy Campaign to our suite of tools, such as fact checking and labels, to help combat climate misinformation and created a new page that explains our holistic approach to addressing climate content and misinformation on our apps.



- The Climate Science Center and it's now available in 165 countries, including New Zealand. In addition, we launched the Climate InfoFinder tool that enables people to search for trusted information about climate change and link to this content directly in comment threads. Finally, we worked with partners to help launch our first Climate Science Literacy Initiative. Its goal is to pre-bunk climate misinformation by running ads across our products and apps that feature five of the most common techniques used to misrepresent climate change.

New Zealand misinformation and media literacy roundtable

- In February 2023, we convened and hosted a workshop and information session in Wellington with the Department of Internal Affairs, NetSafe, the Classifications Office, New Zealand Police, the Electoral Commission, Department of Prime Minister and Cabinet, the Iwi Communications Collective and leading New Zealand academics. Meta facilitated RMIT Cross Check's, Dr. Anne Kruger and the National Association of Media Literacy and Education (NAMLE) to travel to Wellington and host/present this session. It focused on explaining the taxonomy of mis- and disinformation, the drivers, stages and origins, strategies for community engagement and cross sector collaboration. This was the first time such a group had been convened in New Zealand.

- As noted above, in 2023 we partnered with NetSafe to deliver a suite of Micro-learning modules direct to classrooms in New Zealand. One of these modules focused on Media Literacy, developed by New Zealand's Dr. Helen Sissons (an expert in misinformation and media literacy) and Dr. Anne Kruger (above). This module is a practical training tool designed to engage students, educators and their families. It is presently the only course in the New Zealand classrooms focused on media literacy in the digital environment.

Connect, Alert, Inform

- The Pacific is in the most disaster prone region on earth and we know the importance of our services in connecting people before, during and after a natural disaster. In the aftermath of Cyclone Gabriel for example, we saw our services used to coordinate emergency responses and fundraise by government, community groups, NGOs and iwi across the country. We also know that natural disaster events are vulnerable to misinformation. In 2023 we developed a bespoke training programme (Connect, Alert, Inform) for iwi, government, NGOs and community groups across New Zealand and the Pacific attended by over 200 people. The programme included specific modules on managing mis and disinformation in crises delivered by Dr. Anne Kruger of RMIT and Dr. Kate Delmo - a disaster communications academic at the University of Technology in Sydney.

Globally in the first and second quarter 2022:

- For the first quarter, we removed more than 1.7 million pieces of content for violating our COVID-19 misinformation policies across Facebook and Instagram. We displayed warnings on over 180 million distinct pieces of content on Facebook (including reshares) globally based on over 120 thousand debunking articles written by our fact checking partners.
- For the second quarter, we removed more than 1.1 million pieces of content for violating our COVID-19 misinformation policies across Facebook and Instagram. We displayed warnings on over 200 million distinct pieces of content on Facebook (including reshares) globally based on over 130 thousand debunking articles written by our fact checking partners.

For New Zealand, in 2022:

- **We removed over 17,000 pieces of content on Facebook and Instagram in New Zealand violating our harmful health misinformation policy**.
- **We displayed warning labels on over 2 million distinct pieces of content on Facebook in New Zealand (including reshares) based on over 184,000 articles** written by our global third-party fact checking partners.

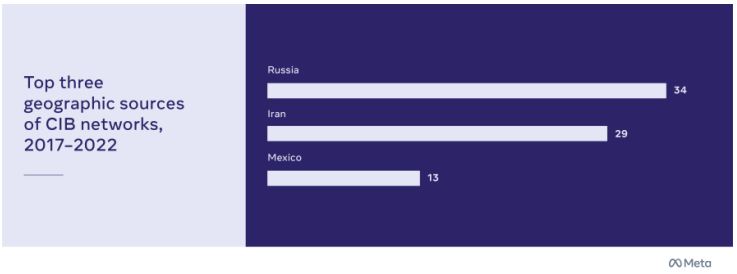
Outcome 7: Provide safeguards to reduce the risk of harm arising from online disinformation

At Meta, disinformation refers to coordinated efforts to manipulate public debate for a strategic goal, with the intention to deceive, and involve *behaviour* that is inauthentic. This is distinctly different from misinformation, which is *content* that is false or misleading.

As outlined in the Baseline Report, we take a three-prong approach to tackling disinformation — 1) preventing interference, 2) fighting misinformation, 3) increasing transparency. We have a team of over 200 experts across the company — with backgrounds in law enforcement, national security, investigative journalism, cybersecurity, law, and engineering — working to disrupt networks of threat actors. We also continue improving our scaled solutions to help detect and prevent the proliferation of inauthentic accounts and behaviours, and partner with civil society, researchers and governments to strengthen our defences.

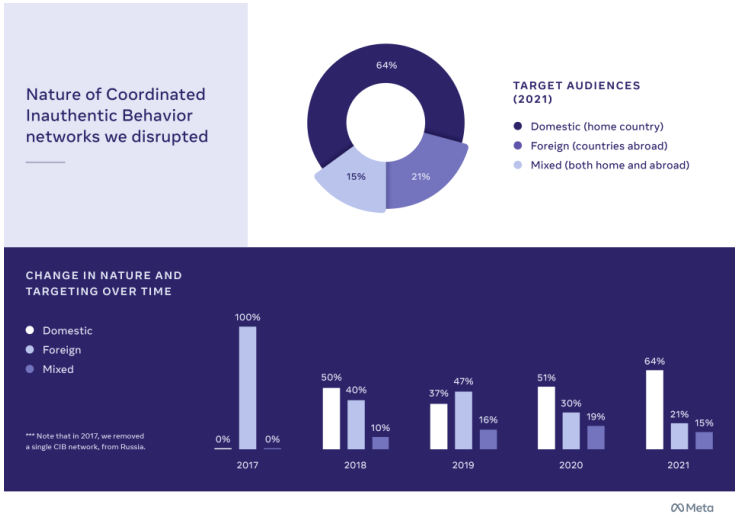
- **Removing networks of Accounts, Pages and Groups that violate our Inauthentic Behaviour policy, including disinformation.** We have maintained the approach outlined in our Baseline Report and have released new data on Meta's CIB take downs. The following are key data insights from our efforts to tackle CIB, as of December 2022. The full set of data can be found here.
    - Since 2017, we have taken down and reported on more than 200 covert influence operations.
    - These networks came from 68 countries and operated in at least 42 languages, with most targeting audiences in their home countries and only around one-third aimed solely at audiences abroad.
    - More than 100 different countries, from Afghanistan to Zimbabwe, have been targeted by at least one CIB network — foreign or domestic — since we began our public reporting. The United States was the most targeted country, followed by Ukraine and the United Kingdom.

    - Russia (34 networks), Iran (29 networks) and Mexico (13 networks) were the three most prolific geographic sources of CIB activity — whether by state actors, political groups or commercial firms.

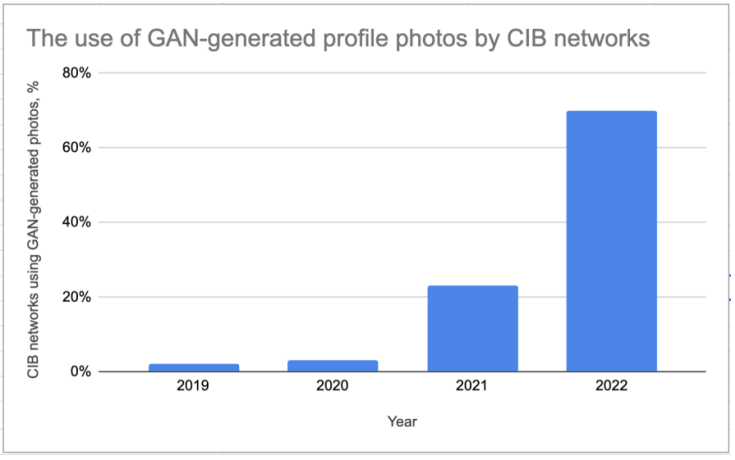


    - CIB networks worldwide most frequently targeted people in their own country. Two-thirds of the operations we've disrupted since 2017 focused wholly or partially

on domestic audiences. Around 90% of CIB operations in Asia-Pacific, sub-Saharan Africa and Latin America were wholly or partly focused on domestic audiences. By contrast, over two-thirds of CIB networks that originated in Europe and the Middle East and North Africa (MENA) were wholly or partly focused on foreign audiences.



- Since 2019, we have seen a rapid rise in the number of networks that use profile photos generated using artificial intelligence techniques like generative adversarial networks (GAN). This technology is readily available on the internet, allowing anyone — including threat actors — to create a unique photo. More than two-thirds of all the CIB networks we disrupted this year featured accounts that likely had GAN-generated profile pictures.



- **Removing fake accounts to keep threat actors off our platforms.** We remove as many fake accounts as we can to minimise opportunities for threat actors to operate on our platforms. These include accounts created with malicious intent to violate our policies; accounts used in spam campaigns and are financially motivated; and benign accounts such as personal profiles created to represent a business, organisation or non-human entity, such as a pet. As reported in our quarterly Community Standards Enforcement Report , we have removed billions of fake accounts. The table below shows the number of accounts removed globally

in 2022 and the proactive rate of fake accounts detected and actioned on before people reported them.

Period	Facebook	Instagram
Jan 2022 - Mar 2022	1.6 billion with proactive rate over 99%	not available
Apr 2022 - Jun 2022	1.4 billion with proactive rate over 99%	not available
Jul 2022 - Sep 2022	1.5 billion with proactive rate over 99%	not available
Oct 2022- Dec 2022	1.3 billion with proactive rate over 99%	not available

- **Raising awareness and providing transparency into our CIB network disruptions.** As noted in the Baseline Report, publishing our research and information about our CIB network disruptions is an important part of our strategy to raise awareness of influence operations threats on our platforms and show the progress we are making. These reports allow researchers, journalists, policymakers, and security experts to scrutinise our work. The quarterly Adversarial Threat Reports and information on specific CIB network disruptions can be found here.
- **Removing Coordinated Inauthentic Behavior From China and Russia.** More information on these network disruptions can be found in our.
    - **China:** In September last year, we took down two unconnected networks in China and Russia for violating our CIB policy. The Chinese-origin influence operation ran across multiple social media platforms, and was the first one to target US domestic politics ahead of the 2022 midterms and Czechia's foreign policy toward China and Ukraine (see the newsroom post for more details). More recently in August 2023, we took down thousands of accounts and Pages that were part of the largest known cross-platform covert influence operation in the world. It targeted more than 50 apps, including Facebook, Instagram, X (formerly Twitter), YouTube, TikTok, Reddit, Pinterest, Medium, Blogspot, LiveJournal, VKontakte, Vimeo, and dozens of smaller platforms and forums. For the first time, we were able to tie this activity together to confirm it was part of one operation known in the security community as Spamouflage and link it to individuals associated with Chinese law enforcement (see the Second Quarter Adversarial Threat Report for more information on this network).
    - **Russia:** Alongside the Chinese network disruption in September last year, we also took down a Russian network — the largest of its kind we've disrupted since the war in Ukraine began — targeting primarily Germany, France, Italy, Ukraine and the UK with narratives focused on the war and its impact through a sprawling network of over 60 websites impersonating legitimate news organisations (see the newsroom post for more details). More recently in August 2023, we blocked thousands of malicious website domains as well as attempts to run fake accounts and Pages on our platforms connected to the Russian operation known as Doppelganger that we first disrupted a year ago. This operation was focused on mimicking websites of mainstream news outlets and government entities to post fake articles aimed at weakening support for Ukraine. It has now expanded beyond initially targeting

France, Germany and Ukraine to also include the US and Israel. This is the largest and the most aggressively-persistent Russian-origin operation we've taken down since 2017 (see the Second Quarter Adversarial Threat Report for more information on this network).

Preparing for the New Zealand 2023 General Election.

- Since early 2023, we have worked with the New Zealand Electoral Commission to ensure they are best placed to use our platforms to provide authoritative information concerning the election and that they have rapid escalation channels to report content of concern to us.

- We have increased our New Zealand based fact-checking partnership over the period of the election with the Australian Associated Press. Although politicians and political candidates continue to be exempt from our fact-checking programme, our fact-checking partners at AAP and AFP will independently select and review claims on Facebook and Instagram. Where they rate something as false, we label it as such and reduce its distribution. They can also receive referrals directly from the public at AFP: https://factcheck.afp.com/contact and AAP: https://www.aap.com.au/make-a-submission/.

- We are also funding the AAP to run a misinformation awareness and education campaign across our platforms during the lead up to both the New Zealand election and the Australian Voice to Parliament referendum.

- We are working with Dr. Anne Kruger at CrossCheck based at RMIT to provide information and education on the taxonomy of misinformation, how it spreads, strategies for combating it and how to report on it.

- The New Zealand Advertising Standards Authority is onboarded to a special reporting channel whereby they can rapidly report paid content of concern to us (in addition to the other reporting channels with NZ Electoral Commission, NZ Police and NetSafe).
- To combat influence operations our specialised global teams, including New Zealand focused staff, will identify and take action against threats to elections, including signs of coordinated inauthentic behaviour across our apps. We engage with all relevant New Zealand agencies, law enforcement and the Electoral Commission on this.
- To protect MPs andcandidates we've extended our Facebook Protect security program to New Zealand. The program has been rolled out to those who might be at a higher risk of being targeted online, such as candidates and public officials, encouraging them to adopt stronger account security protections. We also ran training sessions with all political parties on our policies and tools and how to keep safe during the campaign period.
- When it comes to political advertising we require advertisers running social issue, electoral and political ads to complete our authorisation process and include "Paid for by" disclaimers, and we store these ads in our Ad Library for seven years. This helps ensure that ads on the platform are authentic and transparent. In addition to the transparency information already available in our Ad Library – including the ad creative, who paid for the ad, and who that ad reached – in July 2022 we began including new information about the targeting selections made by advertisers for all of their ads about social issues, elections

and politics. These insights are aggregated at the Page-level across the following categories:

Definitions and Examples of Information Displayed in Ad Library

Category	Definition	Example
Location	Reach people based on locations such as country, region, or city. **Note:** If an advertiser chooses to target an exact location, such as an address or a pin drop radius location, we'll aggregate this targeting data to the nearest city when available and list the targeting group as "Locations in City" such as "Locations in San Francisco."	"100% of the amount spent was targeted to the United States, which includes 2,450 ads about social issues, elections or politics."
Age	Age range with the minimum and maximum ages of people who will see the ads.	"35% of the amount spent was targeted to people aged 55-years-old; 20% of the amount spent was targeted to people aged 21-years old."
Gender	Whether the advertiser selected Men or Women or All (which includes people who have selected any gender or haven't specified a gender).	"75% of the amount spent was targeted to women."
Detailed targeting	Audiences refined by information like demographics, interests and behaviors.	"30% of the amount spent, or 280 ads, were targeted to people interested in sustainability."
Language	Advertisers can choose to show ads to people who use a specific language. Alternatively, they can target all languages.	"90% of the amount spent was targeted to English speakers."
Custom audiences (if they were included or excluded from targeting)	People included or excluded from an audience using sources like website and app traffic. **Note:** This category will also show how frequently advertisers chose to include or exclude Customer Lists. We'll specifically share if a custom audience was derived from a customer list (customer list custom audiences) and provide percentage of spend, ad count, and whether it was included or excluded.	"35% of the amount spent included a custom audience."
Lookalike audiences (if they were included or excluded from targeting)	People included or excluded from an audience with characteristics that are similar to a known audience.	"10% of the amount spent included a lookalike audience."

- By making advertiser targeting criteria available for analysis and reporting on ads run about social issues, elections and politics, we hope to help people better understand the practices used to reach potential voters on our technologies. More information about our political advertising ads transparency tools and policy can be found here.

4.2 Empower users to have more control and make informed choices

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

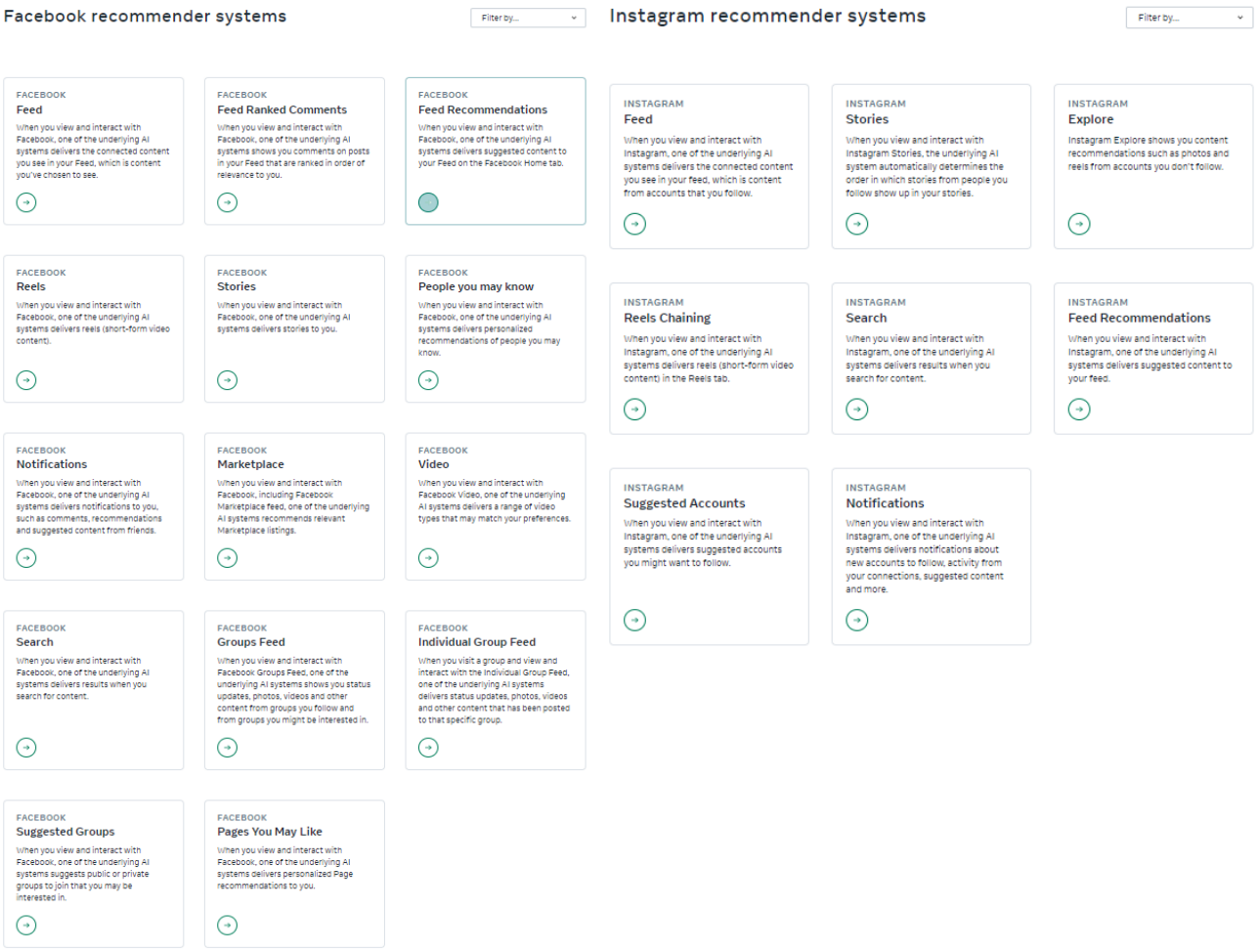
Outcome 9. Users are empowered with control over the content they see and/or their experiences and interactions online

The most effective way to address the online safety and harmful content issue is to build a resilient digital society by providing the tools and resources that will empower people to critically decide for themselves what to read, trust, and share. We do this by providing greater transparency and control to users; providing information that will help them make informed decisions; and advancing media and digital literacy.

- As outlined in the Baseline Report, we offer many tools, products and resources to users to address different areas of safety risks and harms, including:
    - Authoritative information sources
    - Safety hubs
    - Warning labels and notices
    - Parental supervision and age-appropriate controls
    - Comments filtering tools
    - Context buttons with more information
    - Privacy tools
    - Controls to customise what users see in their Feed
    - Feed options that allows users to decide how they want content ranked

AI systems transparency.

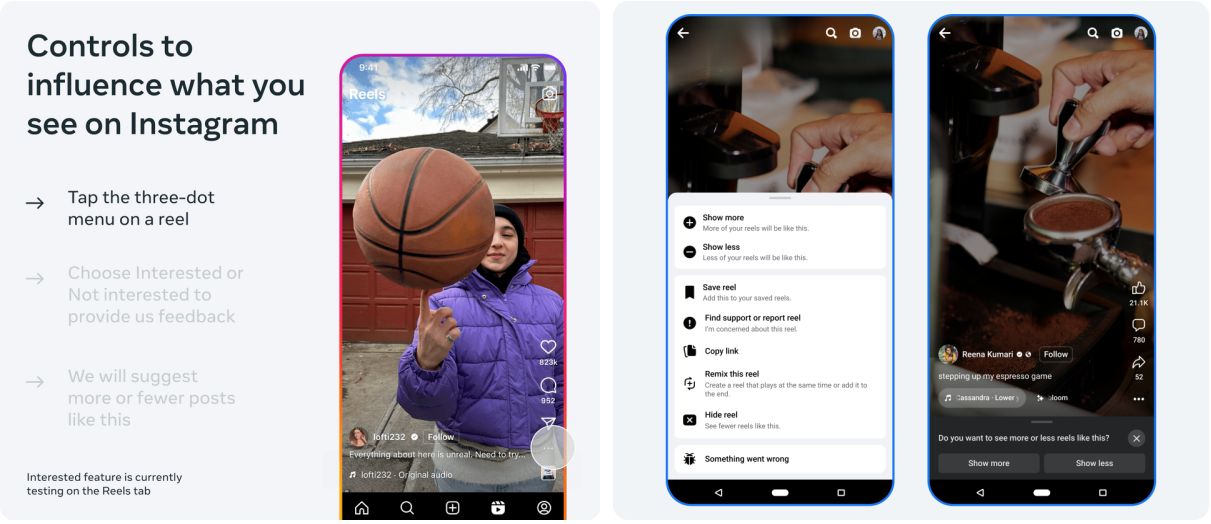
- Since 2014, we have introduced a variety of tools, products and resources to provide greater transparency of our content ranking algorithms (see our Baseline Report). We continue to expand these efforts and in June 2023, we introduced 22 AI Systems Cards for Facebook and Instagram. These provide an in-depth view into how our algorithms work in a way that is accessible for those who do not have deep technical knowledge. They give information about how our AI systems rank content, some of the predictions each system makes to determine what content might be most relevant to people, as well as the controls people can use to help customise their experience.
- The AI Systems Cards cover Feed, Stories, Reels and other surfaces where people go to find content from the accounts or people they follow. The system cards also cover AI systems that recommend “unconnected” content from people, groups, or accounts they don’t follow. More details on the AI Systems Cards can be found here.



Expanding tools to personalise people's experience.

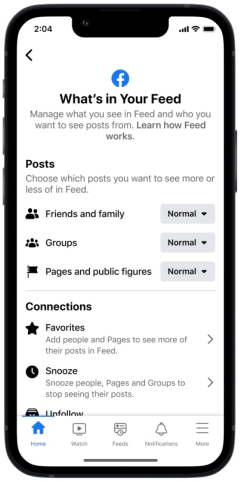
- We have created centralised places on Facebook and Instagram where people can customise controls that influence the content they see on each app. People are now able to visit Feed Preferences on Facebook and the Suggested Content Control Center on Instagram through the three-dot menu on relevant posts, as well as through Settings. On Instagram, we are testing a new feature that makes it possible for people to indicate that they are "Interested" in a recommended reel in the Reels tab, so we can show them more of what they like. The "Not Interested" feature has been available since 2021.
- We also introduced a "Show more, Show less" feature on Facebook, which is available on all posts in Feed, Video, and Reels via the three-dot-menu. On Instagram, people can access "Show more, Show less" by tapping the three-dot menu at the bottom of the video player, as well as for videos in the Watch Feed. If people do not want an algorithmically-ranked Feed – or just want to see what the Feed would look like without it – they can use the Feeds tab on Facebook or select Following on Instagram to switch to a chronological Feed. People can also add other people to their Favourites list on both Facebook and Instagram so they can always see content from their favourite accounts.

  More information about these new tools can be found here.



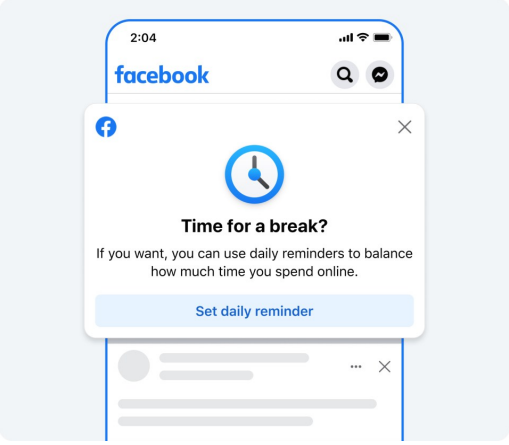
Contextual labels and Feed customisation:

- We also launched new labels on the Reels video player to explain why people see certain reels — for example, because a friend liked it. More details can be found here. Users also have more tools for how much content they see in Feed from friends and family, Groups, Pages and public figures. These tools — as well as Favourites, snooze and reconnect— can be found in Feed Preferences.



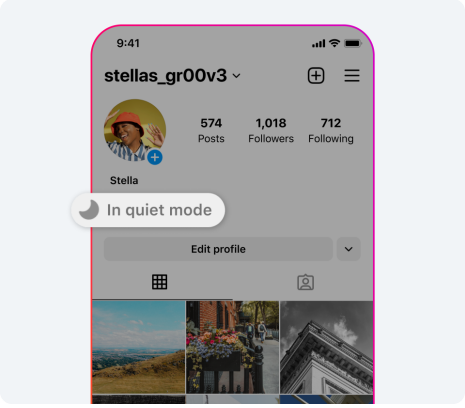
Nudging Teens to Manage Their Time on Facebook and Instagram

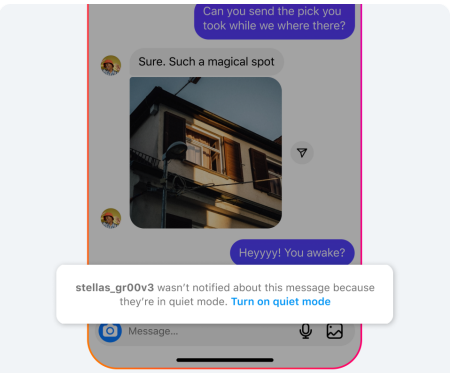
- We've built features like Take a Break on Instagram. Teens see a notification when they've spent 20 minutes on Facebook, prompting them to take time away from the app and set daily time limits. We are now exploring a new nudge on Instagram that suggests teens close the app if they are scrolling Reels at night. More information is available here.

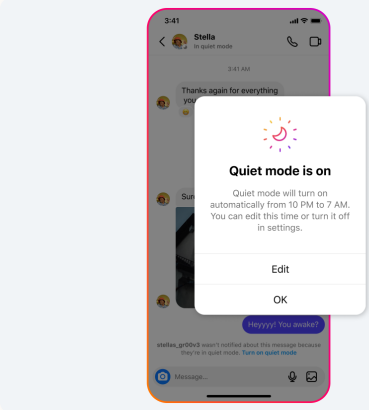


Instagram Quiet Mode: A New Way to Manage Your Time and Focus

- In January 2023, we launched Quiet mode on Instagram to help people focus and to encourage people to set boundaries with their friends and followers. Anyone can use Quiet mode, but we prompt teens to do so when they spend a specific amount of time on Instagram late at night. Quiet mode is available to everyone in the US, United Kingdom, Ireland, Canada, Australia, and New Zealand.

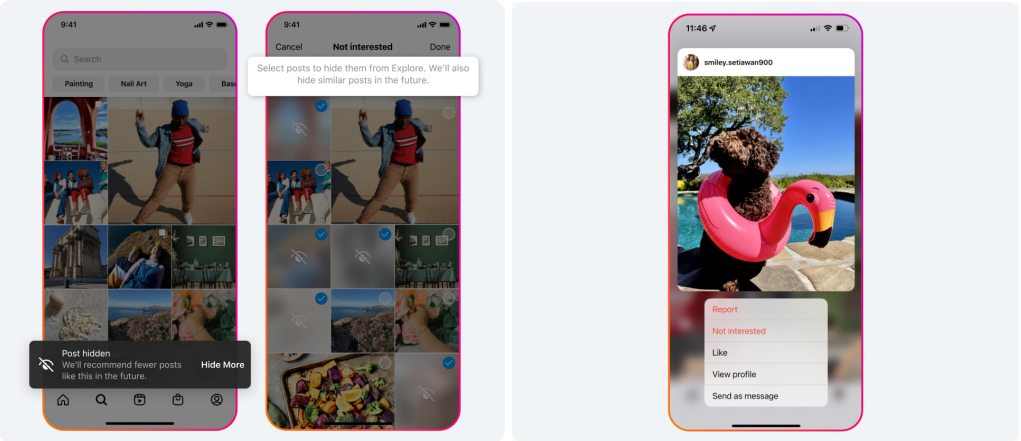




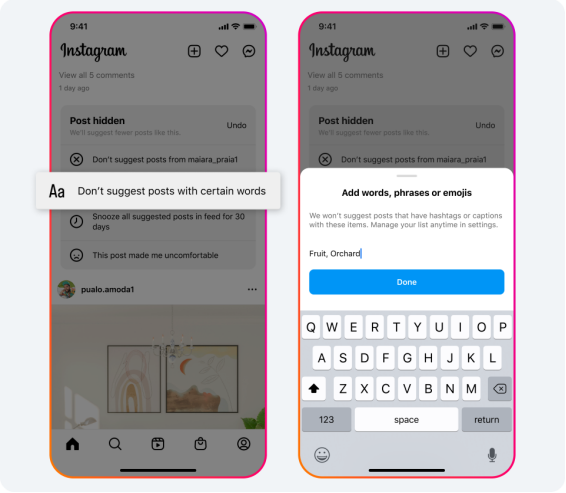


New Ways to Manage Your Recommendations

- We want to give people more control over the content they see on Instagram. Users can now choose to hide multiple pieces of content in Explore that they aren't interested in at one time. They can also select Not interested on a post seen in Explore and we'll avoid showing this kind of content going forward in other places where we make recommendations, like Reels, Search and more.

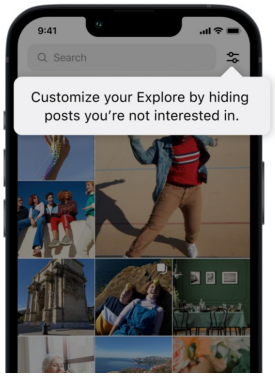


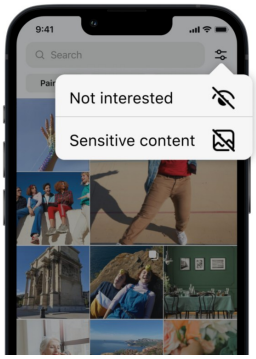
- We have expanded the Hidden Words feature to apply to recommended posts on Instagram. Words, emojis or hashtags users want to avoid — like "fitness" or "recipes" — will no longer be recommend content. Hidden Words are housed in section of Privacy settings.

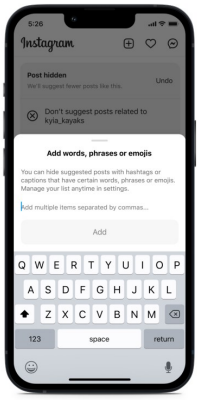


User controls for what they see on Instagram.

- We introduced the ability to mark multiple posts in Explore as Not Interested. We'll immediately hide those posts and refrain from showing you similar content in the future.



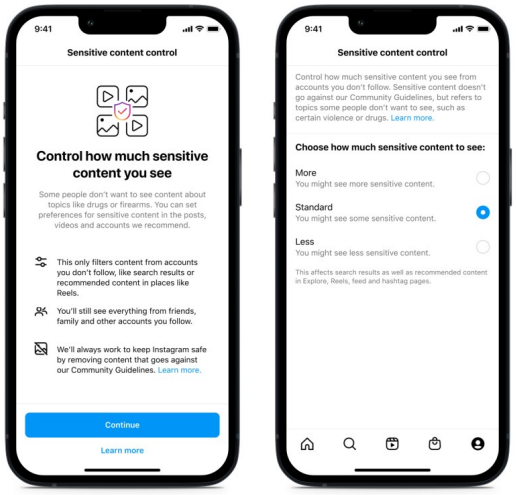




- Users can also now tell Instagram they don't want to see suggested posts with certain words, phrases or emojis in the caption or hashtags, using this feature to stop seeing content that's not interesting to them. Users can also 'Snooze Suggested Posts'

- Users can take a break from suggested posts, by easily 'snoozing' them for 30 days.

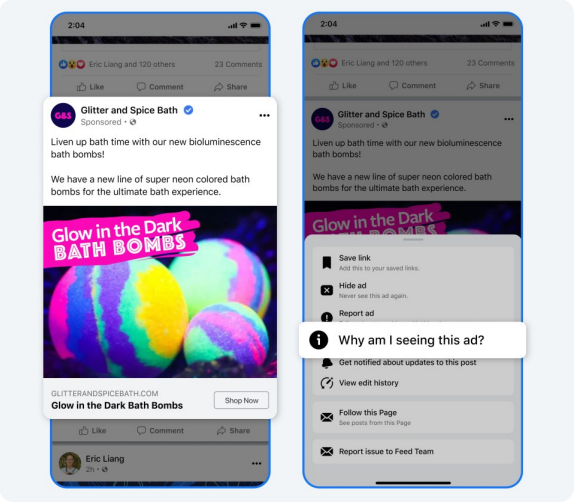
Self Adjusting Sensitive Content Control

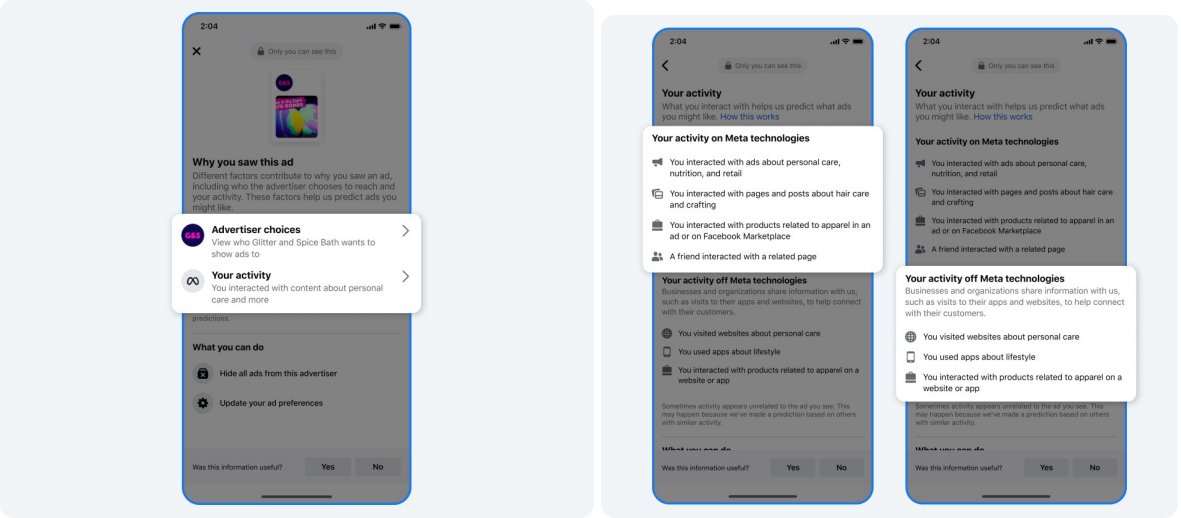


- We've always had rules about what kind of content can be on Instagram, and we call these Community Guidelines. For example, we don't allow hate speech, bullying and other content that might present a risk of harm to people. However, you may see content that doesn't break the rules, but could be upsetting to some.
- We recognize that everybody has different preferences, so you can decide to leave things as they are, or you can adjust the Sensitive Content Control to see more or less of some types of sensitive content. To view your Sensitive Content Control go to your profile, tap the Settings menu, tap Account and tap Sensitive Content control.
- It's important to us that people feel good about the time they spend on Instagram, so we'll continue to work on ways to give people more control over what they see.

Increasing Our Ads Transparency

- In February 2023, we launched the next iteration of the “Why am I seeing this ad?” tool, which we created nearly a decade ago to give people information about why they see certain ads across our technologies. Since its initial launch, we’ve made improvements to “Why am I seeing this ad?” to make it easier to use and understand. We now include information about how we use machine learning models to show people ads. The “Why am I seeing this ad?” tool on Facebook includes:
-
    - Information summarised into topics about how activity both on and off our technologies — such as liking a post on a friend’s Facebook page or interacting with your favourite sports website — may inform the machine learning models we use to shape and deliver ads.
    - New examples and illustrations explaining how our machine learning models connect various topics to show relevant ads.
    - More ways to find our ads controls. Users will now be able to access Ads Preferences from additional pages in the “Why am I seeing this ad?” tool.



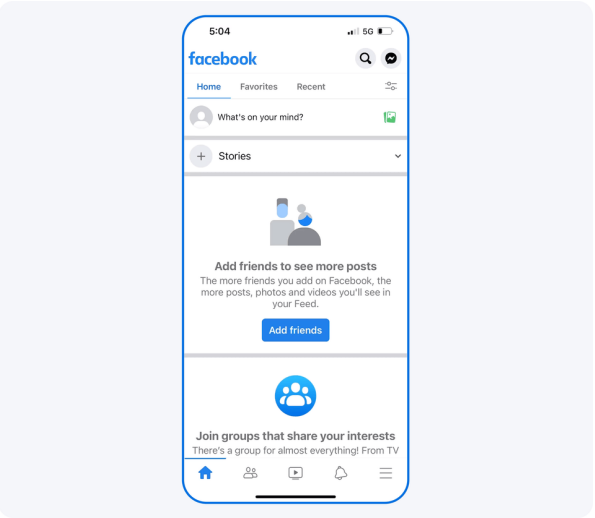


Continuing to Create Age-Appropriate Ad Experiences for Teens

- As part of our continued work to keep our apps age-appropriate for teens, in January 2023 we made further changes to their ad experiences. We recognise that teens aren't necessarily as equipped as adults to make decisions about how their online data is used for advertising, particularly when it comes to showing them products available to purchase. For that reason, we further restricted the options advertisers have to reach teens, as well as the information we use to show ads to teens.
- We also introduced more teen-specific controls and resources to help them understand how ads work and the reasons why they see certain ads on our apps. These changes reflect research, direct feedback from parents and child developmental experts, UN children's rights principles and global regulation.
- The changes we made include:
    - Previously, we made changes to how advertisers can reach teens, which included removing the ability for advertisers to target teens based on their interest and activities. From February 2023, we removed gender as an option for advertisers to reach teens. Additionally, their engagement on our apps — like following certain Instagram posts or Facebook pages — won't inform the types of ads they see.
    - Age and location will be the only information about a teen that we'll use to show them ads. Age and location help us continue to ensure teens see ads that are meant for their age and products and services available where they live.

Giving Teens More Control

- Beginning in March 2023, teens have had more ways to manage the types of ads they see on Facebook and Instagram with Ad Topic Controls, expanding on what's already available. Teens can go to their Ad Preferences within Settings on both apps, and choose See Less or No Preference to further control the types of ads they see.



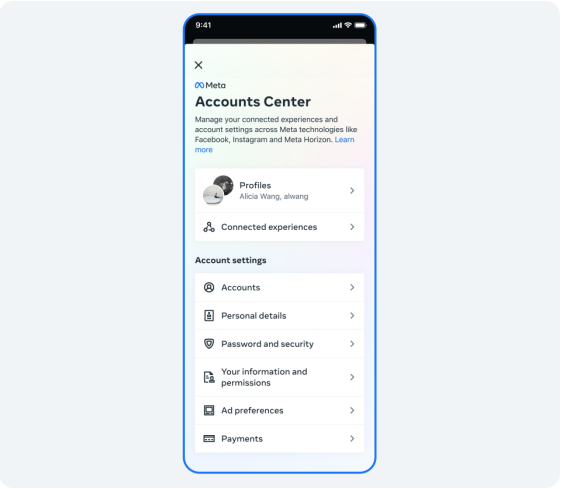
- Our Advertising Standards already prohibit ads about restricted topics — like alcohol, financial products and weight loss products and services — to be shown to people under 18 (and older in certain countries). But even when an ad complies with our policies, teens may want to see fewer ads like it. For example, if a teen wants to see fewer ads about a genre of TV show or an upcoming sports season, they should be able to tell us that.
- Teens can continue to choose to hide any or all ads from a specific advertiser. The topics we already restrict in our policies will be defaulted to See Less, so that teens can't choose to opt into content that may not be age-appropriate.

Helping Teens Understand Their Privacy Options

- We've added a new privacy page with more information for teens about the tools and privacy settings they can use across our technologies, and our teen privacy center has additional resources to help teens understand and manage their privacy across our apps.

Centralising Apps Settings in Accounts Center

- In January 2023, we announced ways to make finding and managing certain settings easier across multiple apps. These improvements can all be found in Accounts Center.



- Personal details, Passwords and security, and Ad preferences now live in a centralised place, so it'll be easier for people who use multiple apps to manage their settings.
- We've also now updated our Data about activity from Partners' control, which is now called Activity information from ad partners to help people easily understand how their activity sent from other websites and apps is used to power the ads they see.

4.3 Enhance transparency of policies, processes and systems

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

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

Meta is committed to making transparent our safety and integrity-related policies, processes and systems where it does not pose a safety and security risk. We believe transparency helps facilitate accountability by making platforms' efforts subject to public scrutiny and, in turn, holds us to account for the decisions we make.

In our Baseline report, we laid out our general views and approach on transparency in section 2, and we have detailed our policies, processes (enforcement), tools and products (systems) in relation to the seven safety and harms themes in section 3. Information on our policies, processes and systems can be found in either our Transparency Center, Help Centers (Facebook, Instagram) or Newsroom.

We have been publishing transparency reports since 2013 because we strive to be open and proactive in the way we safeguard users' safety, security, privacy, and access to information online. We have expanded our reports over the years to include the volume of content restrictions based on local law, the number of global internet disruptions that limit access to our products, reports of intellectual property infringement, and enforcement of our Community Standards/Guidelines. Additionally, we publish reports on our investigations, as well as assessments and evaluations undertaken by Meta or external auditors/consultants, such as the Human Rights report. In addition

to the local New Zealand metrics provided in this Code report, we have also included a summary of the reports we have produced globally, below:

- **Community Standards Report.** We publish the Community Standards Enforcement Report on a quarterly basis to more effectively track our progress and demonstrate our continued commitment to making Facebook and Instagram safe and inclusive.
- **Content restrictions.** We receive reports on content from governments and courts, as well from non-government entities. When content is reported as violating local law, but doesn't go against our Community Standards, we may limit access to that content in the country where the local violation is alleged. This report details instances where we limited access to content based on local law.
- **Government requests for user data.** Meta responds to government requests for data in accordance with applicable law and our terms of service. Each request we receive is carefully reviewed for legal sufficiency and sufficient detail. Meta regularly produces this report on government requests for user data to provide information on the nature and extent of these requests and the strict policies and processes we have in place to handle them.
- **Internet disruptions.** We oppose shutdowns, throttling and other disruptions of internet connectivity and are deeply concerned by the trend towards this approach in some countries. Even temporary disruptions of internet services can undermine human rights and economic activity. That's why we report the number of deliberate internet disruptions caused by governments around the world that impact the availability of our products.
- **Intellectual property report.** We are committed to helping people and organisations protect their IP rights. We do not allow people to post content that violates someone else's IP rights. This report details how many reports of IP violations we received and how much content we took down on as a result.
- **Meta's Quarterly Update on the Oversight Board.** We are committed to publishing regular updates to give our community visibility into our responses to the Oversight Board's independent decisions about some of the most difficult content decisions that Meta makes. The quarterly updates provide regular check-ins on the progress of this long-term work and share more about how Meta approaches decisions and recommendations from the board. These updates provide (1) information about cases that Meta has referred to the board and (2) updates on our progress on implementing the board's recommendations.
- **Human Rights Annual Report.** In July 2022, we released our first annual Human Rights Report which details how we're addressing potential human rights concerns stemming from our products, policies or business practices. We have committed to reporting annually on how we are addressing our human rights impacts, including relevant insights arising from human rights due diligence, and the actions we are taking in response. This report is inspired by Principle 15 of the UN Guiding Principles on Business and Human Rights which makes it clear that companies must "know and show" that they respect human rights.
- **Adversarial Threat Report.** We publish a quarterly adversarial threat report that provides insight into the risks we see worldwide and across multiple policy violations. The report marks nearly five years since we began publicly sharing our threat research and analysis into

covert influence operations that we tackle under the Coordinated Inauthentic Behavior (CIB) policy. Since 2021, we've expanded the areas that our threat reporting covers to include cyber espionage, mass reporting, inauthentic amplification, brigading and other malicious behaviours.

4.4 Support independent research and evaluation

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

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

Meta is committed to supporting independent research that will enhance our understanding of the impact platforms like Meta has on society, as well as investing in research on new content moderation and other technologies that may enhance safety and reduce harmful content online. We also commit to supporting independent evaluation of our systems, policies and processes.

This section describes Meta's efforts to support independent research for the purpose of making our platforms safer and more secure for our users. It also outlines our support and efforts relating to independent evaluation.

Independent Research

- The following are some key initiatives we have supported to empower the independent research community and to help us gain a better understanding of what our users want, need and expect.
    - **Social Science Research.** Meta collaborates with academics and independent researchers around the world and works to provide them with the tools and data they need to study Meta's impact on the world, with a focus on elections, democracy, and well-being.
    - **Data for Good.** In 2017, we launched Data for Good with the goal of empowering partners with data to help make progress on major social issues. A number of New Zealand academics from various disciplines are onboarded to our Data for Good programme.
    - **Research Platform for CIB Network Disruptions.** Since 2018, we have been sharing information with independent researchers about our network disruptions relating to Coordinated Inauthentic Behavior (CIB).
    - **Research Grants & Awards.** Every year, we invest in numerous research projects as part of our overall efforts to make the internet and people on our platforms safer and more secure. The effectiveness of these efforts relies strongly on our partnerships with social scientists to conduct foundational and applied research around challenges pertaining to platform governance in domains such as misinformation, hate speech, violence and incitement, and coordinated harm.

- Meta research awards provide support for independent research projects designed to be shared with the larger scientific, policy, and industry communities. These awards are made as unrestricted gifts to allow investigators the freedom to deepen and extend their existing research portfolios to study the social impact of online interaction and information technologies. The following are some of the key research grants and awards we have supported in the last year:
    - Analysing perceptions of hateful speech using conjoint experiments—Thomas Ralph Davidson (Rutgers University–New Brunswick)
    - Combating Facebook misinformation with local knowledge and community—Susan Fussell, Sharifa Sultana (Cornell University)
    - Countering misinformation in the Southern Hemisphere: A comparative study—Michelle Riedlinger, Silvia Montaña-Niño (Queensland University of Technology), Marina Joubert (Stellenbosch University), Víctor García-Perdomo (Universidad de La Sabana)
    - Digital literacy, demographics, and disinformation—Julia Bernd (International Computer Science Institute)
    - Emotionally-driven, memetic anti-propaganda campaigns—Marina Kogan (University of Utah)
    - Gamifying media literacy interventions for low digital literacy populations—Ayesha Ali, Agha Ali Raza, Ihsan Ayyub Qazi (Lahore University of Management Sciences)
    - Identifying macro and micro factors in spreading conspiratorial content
    - Yu-Ru Lin, Amin Rahimian (University of Pittsburgh)
    - Improving user discernment against inauthentic social media accounts—Mohsen Mosleh (University of Exeter), Cameron Martel, David G. Rand (Massachusetts Institute of Technology)
    - Investigating persuasiveness of contextualised disinformation across media—Aske Mottelson (IT University of Copenhagen)
    - Testing empathy and cognition to reduce later harms from misinformation—Jean Decety, Michael Cohen (University of Chicago)
    - Testing interventions to counter the spread of misinformation—Chico Camargo (University of Exeter)
- **Providing Better Tools for Researchers**
    - We believe an open approach to research and innovation – especially when it comes to transformative AI technologies – is better than leaving the know-how in the hands of a small number of big tech companies. That's why we've released over 1,000 AI models, libraries and data sets for researchers over the last decade so they can benefit from our computing power and pursue research openly and safely. It is our

ambition to continue to be transparent as we make more AI models openly available in future.

- In mid-2023, we began rolling out a new suite of tools for researchers: Meta Content Library and API. The Library includes data from public posts, pages, groups, and events on Facebook. For Instagram, it will include public posts and data from creator and business accounts. Data from the Library can be searched, explored, and filtered on a graphical user interface or through a programmatic API. Researchers from qualified academic and research institutions pursuing scientific or public interest research topics will be able to apply for access to these tools through partners with deep expertise in secure data sharing for research, starting with the University of Michigan's Inter-university Consortium for Political and Social Research. These tools will provide the most comprehensive access to publicly-available content across Facebook and Instagram of any research tool we have built to date and also help us meet new data-sharing and transparency compliance obligations.

Independent Evaluation

We believe independent evaluation is important to hold companies like Meta accountable and help us do better. In addition to this Code, we have participated in several other voluntary initiatives to strengthen accountability of platforms through increased transparency and independent evaluation, e.g. the EU Code of Practice on Disinformation, the Australian Code of Practice on Disinformation and Misinformation and the Digital Trust and Safety Partnership (DTSP).

- In July 2022 the DTSP unveiled its inaugural evaluation of how ten leading technology companies–including Meta–are adhering to Trust & Safety best practices. The report highlighted successes among DTSP partners, including how Trust & Safety teams and functions have performed relatively well when it comes to core practices and activities that fall squarely within their domain and can be implemented unilaterally, to some degree. These include constituting the teams responsible for content policies and developing public facing policy descriptions, as well as developing enforcement infrastructures that span people, processes, and technology, and notifying users whose content is subject to an enforcement action for violating policies. Furthermore, the report also identified areas for improvement. Three of the practices deemed least mature, according to the self-assessments, related to incorporating user and third-party perspectives into Trust & Safety policy and practices. This illustrates the internal focus of Trust & Safety functions. As a discipline, Trust & Safety has developed with less external engagement outside of companies until recently. The least mature of all assessed practices is the creation of processes to support academic and other researchers working on relevant subject matter.

- We have also subjected our content moderation practices to independent assessments and audits by experts, namely the Data Transparency Advisory Group (DTAG) which published an assessment in 2019 on our effectiveness in enforcing our Community Standards, and EY who published an audit report on the accuracy of our metrics in the Community Standards Enforcement Report.

- In May 2023, the World Economic Forum - Digital Safety Coalition released the Digital Safety Risk Assessment in Action, which presents a blueprint for understanding and assessing digital safety risks.

- The report outlines a proposed “industry standard” risk assessment framework accompanied by a bank of case studies demonstrating how the framework might be applied in practice.
- The risk assessment framework draws on existing human rights frameworks, enterprise risk management best practices and evolving regulatory requirements to clarify the factors that should be used to clarify digital safety risks and sets out a methodology for how stakeholders should assess these risk factors in the digital ecosystem.
- The case studies highlight the variety and interconnectedness of existing risk assessment frameworks and approaches while substantiating the complexity of the subject matter, providing an overview of how existing frameworks are designed and leveraged and how a risk assessment framework can be applied in practice to a specific technology, type of harm or type of service