Meta's baseline report for the Aotearoa New Zealand Code of Practice for Online Safety and Harms

November 2022

Meta

Meta's baseline report for the Aotearoa New Zealand Code of Practice for Online Safety and Harms | November 2022

Table of Contents

1. 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"). We believe the Code is a credible step in encouraging collaboration between the technology industry, civil society and governments to combat online harms.

We are looking forward to further enhancing a New Zealand perspective on policy issues through the processes established in this Code, including to inform our own policy development processes, research areas, programmatic initiatives and transparency.

We recognise Meta's responsibility to protect the safety of people who use our services. It is inherent and essential to our business: New Zealanders, other people around the world and businesses globally will only continue to use our platform if they have positive, meaningful and safe experiences.

We have made significant investments in safety. Globally we have more than 40,000 people working on safety and security. We've invested more than US$13 billion on safety and security since 2016, and we spent more than US$5 billion in 2021.

This report gives an overview of the various policies, enforcement techniques, tools, products, resources and partnerships we have developed to enhance the safety and security of our users in relation to our commitments under the Code. The aim of this report is to provide **baseline information** to increase understanding on how Meta approaches online safety and harms. We believe that transparency of such information will help inform the public discourse and policy development.

Meta has opted into all four commitments (a total of 13 outcomes with 45 measures) under the Code. A copy of our participation form can be found here.

For the first time, due to our commitments under this Code, we are also providing **New Zealand specific metrics** for a number of harm categories on content created in New Zealand that we have taken action on. The metrics represented are for the period of January to December 2021, unless otherwise specified, and can be found in the transparency section of each harm category. In addition, we are providing details of the New Zealand specific programmatic activity, research and partnerships we undertake to further localise and inform our global efforts on online safety.

2. Meta's approach to online safety and harm

Meta's approach to online safety 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.

Policies

Meta maintains a robust set of policies (i.e. Community Standards; Community Guidelines; Advertising Standards; Content Distribution Guidelines, Privacy Policy, among others) that allow us to take action on content and accounts, which are central to our approach to reducing the spread of harmful content. The following are our most commonly referenced policies.

- Meta's **Community Standards**[1] (Facebook) and **Community Guidelines**[2] (Instagram) govern what is allowed or not allowed on our platforms. Safety is a core value in the standards, alongside privacy, authenticity, voice, and dignity.[3] These standards prohibit or restrict various categories of online harms, including the seven identified in the Code - 1) child sexual exploitation and abuse; 2) bullying or harassment; 3) hate speech; 4) incitement of violence; 5) violent or graphic content; 6) misinformation; and 7) disinformation. These standards may go further than what is prohibited by law in some instances, as they aim to mitigate the risk of speech content that may lead to real world harm.
- **Repeat violators** (i.e. people that repeatedly violate our Community Standards) may in addition to having their content removed receive decreased distribution, be limited on their ability to advertise or monetise, be blocked from posting new content, or removed from our platforms altogether. Information on how we enforce our standards against repeat violators can be found here.
- Our **Content Distribution Guidelines** details our policy for content that receives reduced distribution, including posts that have been rated false by third-party fact-checkers, are borderline or potentially policy-violating, are low quality or sensationalist, or are comments that are likely to be reported or hidden. The full list of content we demote can be found here.[4]
- Our **Advertising Standards** provide policy detail and guidance on the types of ad content we allow and prohibit. This includes our policy on **social issue, electoral or political ads**. Our Advertising Standards also provide guidance on advertiser behaviour that may result in restrictions being placed on a business account or its assets (i.e. an ad account, Page or user account). The full set of advertising policies can be found here.[5]

We take great care to craft policies that are inclusive of different views and beliefs. Our policies are constantly being updated to keep pace with changes happening online and offline around the world. We run a regular meeting called the Policy Forum[6] to discuss potential changes to our standards. The Forum helps to factor in cultural differences on

[1] https://www.facebook.com/communitystandards
[2] https://help.instagram.com/477434105621119/
[3] https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/
[4] https://transparency.fb.com/features/approach-to-ranking/types-of-content-we-demote
[5] https://transparency.fb.com/en-gb/policies/ad-standards
[6] https://transparency.fb.com/en-gb/policies/improving/policy-forum-minutes/

what is acceptable and better understand broad perspectives on safety and voice and the impact of our policies on communities globally. Updates to our policies are tracked in the change log so people can see how they have evolved.

These policies are developed based on feedback from our community, and the advice of experts in fields such as public, women, child and youth safety; human rights; cybersecurity; and technology. We also seek input from our global network of 400+ safety partners and our Safety Advisory Board. A variety of other internal and external stakeholders are also consulted for local/regional and functional expertise.

Enforcement

Our policies allow us to enforce against a broad range of violating activity across three specific areas:

1) **Actor-based enforcement**, which involves the removal of accounts or organisations because of the totality of their activity on the platform;

2) **Behaviour-based enforcement**, which is predicated on specific violating behaviours exhibited by violating actors; and

3) **Content-based enforcement**, which predicates enforcement on specific violations of our Community Standards.

Online safety and integrity issues are a complex problem. The public debate often treats this as one problem, but it is a variety of different problems rolled together. When we blur issues together as one problem set, it becomes very hard to develop a strategy to solve any one part. This is why we break this problem down into the three dimensions - actors, behaviours, and content. For example, any potential violation could be conducted by a problematic actor (e.g., a terrorist or criminal organisation); using problematic behaviour (e.g., networks of fake accounts); distributing problematic content (e.g., false information or hate speech). Our policies work along each dimension, which we then tailor our response to the nature of the violation. This gives us a range of tools to respond with. Having a coherent and comprehensive approach across all three dimensions provides us with a network of enforcement operations.

Enforcement of our policies will never be perfect given the dynamic nature and scale of online speech, the limits of enforcement technology, and the different expectations people have over their privacy and their experiences online. Meta, and other companies, may be limited in our ability to review speech for allegations of local policy or legal violations, due to a lack of context that is often necessary.

Taking these challenges into consideration, we apply a three-part strategy - **remove, reduce, and inform**[7] - to reduce the prevalence of harmful content and activity across the Meta family of apps, while helping to build a digitally resilient society where people are better able to critically evaluate information, make informed decisions and correct mistakes themselves. This involves removing content and accounts that violate our policies; reducing the spread of problematic content that does not violate our policies but still undermines the authenticity and integrity of the platform, and providing people with more information that will help them make more informed decisions on the content they see.

[7] https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/

- **Remove:** We remove content and accounts that violate our Community Standards & Guidelines, including fake accounts and accounts engaged in inauthentic behaviour, misinformation that is likely to contribute to the risk of imminent physical harm, voter fraud or interference, hate speech, bullying and harassment. We also remove ads that violate our Advertising Standards[8], including ads that violate our Community Standards and ads with debunked claims by third-party fact-checkers or, in certain circumstances, by authoritative bodies.

- **Reduce:** Problematic content that does not meet the standards for removal under but still undermines the authenticity and integrity of the platform, such as clickbait and content debunked by our third-party fact-checkers, are demoted in the Feed. This significantly reduces the number of people on Facebook and Instagram who see that content. Our Content Distribution Guidelines provide transparency about how we define and treat problematic or low-quality content.

- **Inform:** We help prevent the spread of harmful content by providing additional context and connecting people with accurate and authoritative information so that they can make informed decisions about the content they see, post and share.

This section focuses on the first two parts of this strategy - remove and reduce. The third part of our strategy - inform - will be addressed in the sections below on tools, products and resources and user empowerment.

Content Review

We invest significantly in technology and people to detect and identify violating content, accounts, or suspicious behaviour. Our enforcement practices consist of three pillars - **artificial intelligence (AI), human review, and user reports**.

- **Artificial intelligence (AI):** A central focus of Meta's AI efforts is deploying machine learning technology to protect people from harmful content and accounts posting such content. Billions of people use our platforms, so we rely on AI to scale our content review work and automate decisions when possible. Our goal is to spot violating content quickly and accurately before people have the chance to see it.

  We continue to improve our AI systems to proactively detect violations across a wide variety of areas without relying on users to report content to us, often with greater accuracy than reports from users. Last year, we built and deployed a new AI technology, Few-Shot Learner (FSL), that can learn and adapt to take action on new or evolving types of harmful content within *weeks* rather than months. FSL learns from different kinds of data, such as images and text.[9]

  As shown in our quarterly Community Standards Enforcement Reports[10], we increasingly find and take action on violating content and accounts before people report it. For example, from April to June 2022, we found and took action on 95.6% of hate speech content on Facebook before people reported them. Only 4.4% of hate speech were taken down as a result of user reports, compared to 76.4% in the fourth quarter of 2017.

[8] https://transparency.fb.com/policies/ad-standards
[9] https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/
[10] https://transparency.fb.com/data/community-standards-enforcement/

- **Human review:** We have over 40,000 people dedicated to keeping people safe and secure on our platforms. This includes expert teams in safety, cybersecurity, human rights, counterterrorism, social science, and local markets (i.e. regions, countries). Our content reviewers come from many cultural backgrounds, reflect the diversity of our community, and bring a wide array of professional experiences. We also employ market specialists to provide societal and cultural context and additional operational capability to our enforcement systems for the region they focus on and this includes New Zealand and the Pacific Islands.

- **User reports:** As we improve our AI capabilities to proactively detect and take action on violating content, our reliance on user reports have significantly decreased. However, every week, people around the world report millions of pieces of content to us that they believe violate our policies.

  We strive to improve our reporting tools to make it easier for people to report content they think may violate our policies, but there are limitations. For example, in areas with lower digital literacy, people may be less aware of the option to file a user report. People also often report content they may dislike or disagree with, but that does not violate our policies. For example, users may report content from rival sports teams or opinions they do not agree with. In addition, some content may be seen by a lot of people before it is reported, so we can't rely on user reports alone. Information on how to report something can be found here.

Tools, Products and Resources

We believe that users should be empowered to customise their online experience and be given tools and resources to help keep themselves safe on the platform. Our strategy is to design tools that: 1) raise awareness of online safety issues, 2) provide access to accurate and credible information, 3) give more context on posts, and 4) provide people with more control over their online experience. The design and development of these tools are informed by consultations with industry, experts and civil society organisations. We also have a number of tools, products and resources that target specific online harms, including the seven harm categories outlined in the Code. These are outlined in Section 3, under each harm area.

The following are some of the tools, products and resources we have deployed to mitigate the risk of harmful experiences on our platforms in general.

- **Blocking followers.** We provide options to Block, Report, Hide or Unfollow users. To protect users from unwanted contact on Instagram, we launched additional features for users to also block existing and new accounts the originally blocked user creates. This is designed to help make sure users don't hear from people they've blocked, even when they create a new account.

- **Comment Controls.** We provide options for users to decide who is allowed to comment on their public posts, as well as profanity and keyword filters. On Instagram, for example, the Comment Controls feature allows users to automatically hide comments based on a list of words, phrases, numbers or emojis that they determine, based on their experiences or preferences. If people comment using

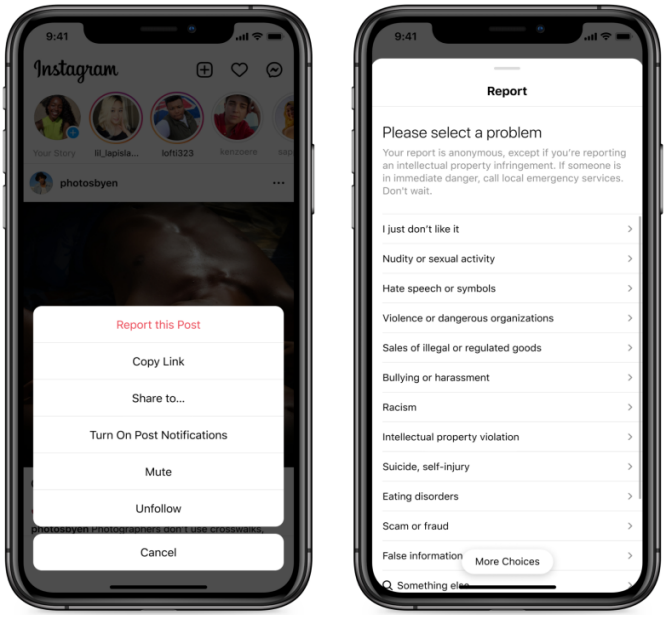
those words or emojis, the user will not be notified and the filter comment will not be published on the post for anyone to see.[11]

- **Access to Authoritative Information and Resources.** Throughout our platform, we make information, tips and resources available at appropriate "just-in time" points. For example, if someone searches for "COVID-19", they are directed to the COVID-19 Information Center. These resources are mostly developed in partnership with safety experts, media/digital literacy organisations, public health authorities, universities and other trusted sources. This includes:
    - Information Centers that provide reliable, up-to-date information from trusted and credible sources on COVID-19[12] and Climate Science[13]. When people search for information on these topics, they are provided links to the Information Centers. We also apply information labels to some posts on these topics that direct users to the Information Centers.
    - Instagram Safety and Wellbeing Hub[14] and the Facebook Safety Center[15] that provide users with tools to stay safe, secure their accounts and protect their information.
    - Youth Portal provides a central place for teens to access education on our tools and products, first person accounts from teens about how they're using technologies, tips on security and reporting, and advice on how to use social media safely.[16]
    - Suicide Prevention Support Center that provides resources and guidance on how to access and offer support.[17]
    - Get Digital Hub, a digital citizenship and wellbeing program which provides schools and families with lesson plans and activities to help build the core competencies and skills young people need to navigate the digital world in safe ways.[18]
    - Digital Literacy Library, developed in partnership with the Berkman Klein Center for Internet & Society at Harvard University, provides lesson plans for parents and educators to help young people develop skills needed to navigate the digital world, critically consume information and responsibly produce and share content. Lessons involve group discussions, activities, quizzes, and games that have been built in consultation with teens.[19]
    - Bullying Prevention Hub, developed in partnership with the Yale Center for Emotional Intelligence, is a resource for educators and families seeking support for issues related to bullying and other conflicts. It offers step-by-step plans, including guidance on how to start important conversations for

[11] https://about.instagram.com/en_US/blog/announcements/national-bullying-prevention-month
[12] https://www.facebook.com/coronavirus_info/
[13] https://www.facebook.com/climatescienceinfo
[14] https://about.instagram.com/safety
[15] https://www.facebook.com/safety/tools
[16] https://www.facebook.com/safety/youth/
[17] https://www.facebook.com/safety/wellbeing/suicideprevention/
[18] https://www.facebook.com/fbgetdigital
[19] https://www.facebook.com/safety/educators

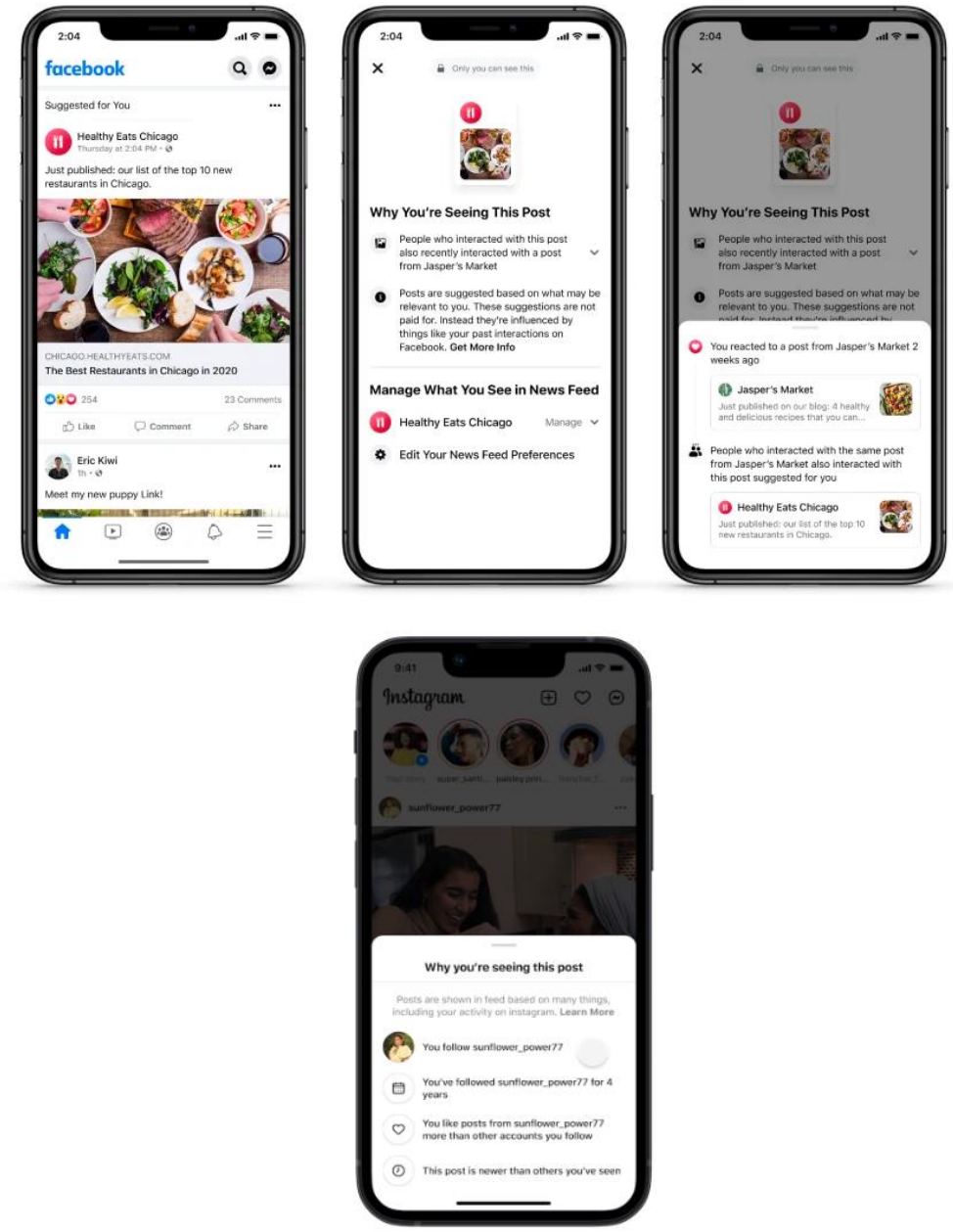
people being bullied, advice for parents and caregivers who have a child that's being bullied or accused of bullying, and educators who have had students involved with bullying.[20]

- **User Reporting.** While we continue to innovate and improve our technologies to combat new trends and techniques that abusive accounts may use, we also allow users to report individual pieces of content or accounts that they believe are violating our policies. As noted in our general approach, user reports is one of the three pillars where we may get signals of violating content or behaviours.



- **Ranking Transparency & Controls.** We want people to better understand why they see certain posts and ads in their Feed and be able to decide if they want to continue seeing those posts or ads. Since 2014, we have introduced tools to explain how people's past interactions impact the ranking of posts in their Feed. A full list of ranking tools, products and resources can be found here.[21]
    - Why am I seeing this post? helps people understand and more easily control what they see from friends, Pages, and Groups in their Feed. It explains how people's activities impact the ranking of content in their Feeds (e.g. if the post is from a friend, Page or Group the user follows), as well as what other information are largely influencing the order of posts (e.g. how often the user interacts with a specific type of post like videos, photos, links). This tool also provides shortcuts to controls, such as See First, Unfollow, Feed Preferences, and Privacy Shortcuts, to help users personalise their Feed.[22]

[20] https://www.facebook.com/safety/bullying
[21] https://about.fb.com/news/2021/12/changes-to-news-feed-in-2021/
[22] https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/



- Why am I seeing this ad? allows people to see how factors like basic demographic details, interests and website visits contribute to the ads that are shown in their Feeds. There are also additional details about when information on an advertiser's audience list matches a person's profile.[23]
- Instagram System Cards[24] help people understand how AI shapes their product experiences and provide insights into how the Feed ranking system dynamically works to deliver a personalised experience on Instagram. Users can test to rank hypothetical users' Feed to see how it compares with what the feed system might predict. Information on the research that led to the development of the System Cards can be found here.[25]

[23] https://about.fb.com/news/2019/07/understand-why-youre-seeing-ads/
[24] https://ai.facebook.com/tools/system-cards/instagram-feed-ranking/
[25] https://ai.facebook.com/research/publications/system-level-transparency-of-machine-learning

- **Feed Ranking Controls**. We have made it easier for people to control what they see in their Feed by offering:
    - Feed Preferences provides a suite of tools that allow people to manage what they see in their Facebook Feed, including the ability to unfollow people, snooze a particular account, or prioritise Favourites.[26]
    - Favourites Feed allows users to control and prioritise posts from friends and Pages they care about most. By selecting up to 30 friends and Pages to include in Favourites, their posts will appear higher in ranked News Feed and can also be viewed in a separate feed populated exclusively with posts from a person's "Favourites" (see screenshot below).[27]
    - Most Recent Feed allows users to see content sorted in their Feed by chronological order with the newest post first.
    - Feed Filter Bar allows users to alternate between different Feed experiences - the algorithmically-ranked Top Posts Feed, the chronological Most Recent Feed[28], or the Favourites Feed.[29]

Partnerships

We have over 400 safety partners around the world, including a number of partnerships in New Zealand, to ensure our global safety efforts are complemented by on-the-ground expertise.

We have a Safety Advisory Board, which comprises leading safety organisations and experts from around the world. Board members provide expertise and perspective that inform Meta's approach to safety. Netsafe – New Zealand's independent, not-for-profit

[26] https://www.facebook.com/help/371675846332829
[27] https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/
[28] https://www.facebook.com/help/218728138156311
[29] https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/

online safety organisation – is one of only 11 organisations globally that serves on this Board.

Additionally, we work with a number of New Zealand organisations to flag harmful content to us. These organisations include NetSafe, the Department of Internal Affairs, the Electoral Commission and New Zealand Police.

We have a long-standing relationship with NetSafe with respect to online safety. The relationship consists of: 1) operations - NetSafe can flag content to us that has been directly reported to them by the public or other groups via a dedicated reporting channel; 2) policy - we collaborate on initiatives that promote a strong online safety ecosystem in New Zealand (including the development of this Code); and 3) education - we partner to develop online safety campaigns and tools to enhance user safety. Over the past 10 years, we have worked, and continue to work, with NetSafe on a variety of issues, including bullying and harassment, child exploitation, domestic abuse, misinformation, suicide and self-injury, non-consensual sharing of intimate images (NCII - also known as 'revenge porn'), scams and many other areas of online harm.

In addition, we partner with Youthline which provides significant mental support for young people in New Zealand. We have supported Youthline for a number of years, including through a combination of grant funding for their operations (especially during the COVID-19 lockdowns), advertising credits (including to promote their service in particular localities following a suicide or self-injury (SSI) incident), and research (a report on social media and postvention support will be released in early 2023).

Concerning misinformation, especially as it relates to COVID-19 and vaccinations, we have partnered worked alongside a number of organisations including NetSafe, Prepare Pacific, the Ministry of Health, Unite Against COVID-19, Ministry for Pacific Peoples, Te Puni Kokiri, the New Zealand Police, AAP and AFP (fact-checkers). We also partner with Cross Check, a First Draft initiative, that focuses on programmes and research to combat misinformation across the media ecosystem.

We are a signatory to the Christchurch Call to Action, contributing to a number of working groups focused on community-building; crisis and incident response; algorithms and positive interventions; and transparency. Separately, we have a partnership with Sakinah Community Trust to promote community cohesion through Unity Week. The Trust is a women-led organisation comprising the next-of-kin of those lost in the Christchurch mosque attacks of 15 March 2019, that focuses on the development of long-term community response and engagement.

Transparency & Accountability

As a large company the decisions we take relating to content can be significant. Accurate and meaningful transparency is critical to holding platforms accountable. No company should mark its own homework, and the credibility of our systems should be earned, not assumed. We therefore support a number of initiatives to ensure there is enhanced transparency and accountability for the decisions we take.

We provide transparency into a wide range of areas – our community standards enforcement, government and law enforcement requests, content restrictions, internet

disruptions – among others. We have also committed to external checks and audits of several of these transparency measures.

- **Transparency Center.** A Meta Transparency Center[30] provides visibility into our extensive policies, how we enforce those policies, respond to data requests from governments and protect intellectual property, while monitoring dynamics that limit access to Meta's platforms.

- **Community Standards Enforcement Report.** We publish the Community Standards Enforcement Report (CSER) on a quarterly basis to track our progress. The report is a voluntary effort that allows for scrutiny of our work to enforce Facebook and Instagram's Community Standards/Guidelines. We continue to expand the CSER, which now reports on five metrics — content removed; content removed proactively; prevalence; appeals; and restored content — across 11 policy areas.

    a. Adult Nudity and Sexual Activity
    b. Bullying and Harassment
    c. Child Endangerment
    d. Dangerous Organisations
    e. Fake Accounts
    f. Hate Speech
    g. Regulated Goods
    h. Spam
    i. Suicide and Self-Injury
    j. Violence and Incitement
    k. Violent and Graphic Content

- **Independent Assessment.** Independent audits and assessments are crucial to hold companies like Meta accountable and help us do better. To that end, we have engaged in two important initiatives:

    1. An independent, public assessment in 2019 by the Data Transparency Advisory Group (DTAG) — a group of international experts in measurement, statistics, law, economics and governance — of whether the metrics we share are useful, based on sound methodology and accurately reflect what's happening on our platform. DTAG broadly agreed that we are looking at the right metrics and provided some recommendations for improvement including launching appeals reporting.[31]

    2. An independent audit by Ernst & Young, to validate whether the CSER is measured and reported correctly. In that audit, EY found the calculation of

[30] https://transparency.fb.com/data/
[31] The full Report of the Facebook Data Transparency Advisory Group can be found here https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf

the metrics in the CSER were accurate, were fairly stated, and that internal controls were suitably designed, and operating effectively.[32]

- **Oversight Board.** We established the Oversight Board, an independent group of experts who review important content decisions we make, and help us balance free speech and safety. We publish quarterly reports to provide information about cases that Meta has referred to the board and updates on our progress in implementing the board's recommendations.

- **Digital Trust & Safety Partnership (DTSP).** We are a founding member of the DTSP, a first-of-its-kind partnership that brings together technology companies of different sizes and business models to tackle online safety concerns. The partnership aims to develop a trust-and-safety focused **best practice framework** that was developed with input and collaboration with consumer/user advocates, policymakers, law enforcement, relevant NGOs and various industry-wide experts. The framework covers the entire product lifecycle, broken down into five areas - development, governance, enforcement, improvement, transparency.[33] Participating companies are required to complete **self-assessments** on the steps taken to identify and address content- and conduct-related risks on their products/services, in relation to the best practice framework, which are subject to **third-party, independent evaluation** (to be completed in 2023).

- **Global Network Initiative**. As set out in our Corporate Human Rights Policy, we're committed to understanding the role our platforms play offline and how our products and policies can evolve to create better outcomes. Engaging independent experts and evaluating our work through the lens of global human rights principles is key to achieving this goal.[34] As a member of the Global Network Initiative (GNI), we are committed to upholding the human rights standards set out in the GNI Principles on Freedom of Expression and Privacy and Implementation Guidelines. We are independently assessed on our implementation of the GNI Principles every three years, in which the results are made public for people to read and scrutinise.[35] In line with our GNI commitments, we also commission third-party human rights impact assessments (HRIA) to help us identify and mitigate potential human rights risks and impacts. Many of the HRIAs are made public.[36]

3. Reducing the prevalence of harmful content online

As required in section 4.1 of the Code, this section outlines the policies, processes, products and programs that Meta deploys to promote online safety and mitigate risks that may arise from the propagation of harmful content online, as it relates to the harm categories identified in the Code:

1. Child sexual exploitation and abuse
2. Cyberbullying or harassment
3. Hate speech

[32] https://about.fb.com/news/2022/05/community-standards-enforcement-report-assessment-results/
[33] https://dtspartnership.org/best-practices/
[34] https://humanrights.fb.com/
[35] https://globalnetworkinitiative.org/
[36] https://humanrights.fb.com/our-impact/

4. Incitement of violence
5. Violent or graphic content
6. Misinformation
7. Disinformation

Our efforts for each theme follow our general five-part approach to online safety and harm — 1) policies, 2) enforcement, 3) tools and products, and resources, 4) partnerships and 5) transparency.

Child Sexual Exploitation and Abuse (CSEA)

Meta takes a comprehensive approach to child safety, including zero-tolerance policies prohibiting child exploitation; technology to prevent, detect, remove and report policy violations; and victim resources and support. We collaborate with industry, child safety experts and civil society around the world to fight the online exploitation of children because our work in this space extends beyond our apps to the broader internet. We are also developing targeted solutions, including new tools and policies to reduce the sharing of child exploitative content on Facebook and Instagram. 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.

Policies

We do not allow content that sexually exploits or endangers children. When we become aware of apparent child sexual exploitation, we report it to National Center for Missing and Exploited Children (NCMEC) in compliance with applicable US law. 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.

Our Child Sexual Exploitation, Abuse and Nudity[37] policy bans content or activity that threatens, depicts, praises, supports, provides instructions for, makes statements of intent, admits participation in or shares links of the sexual exploitation of children (i.e. real or non-real minors, toddlers or babies). To avoid even the potential for abuse, we take action on nonsexual content as well, such as seemingly benign photos (e.g. photos of children in the bath). We know that sometimes people share naked 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 remove accounts that promote this type of content. Specifically, our policy covers:

- Child sexual exploitation
- Solicitation
- Inappropriate interactions with children
- Exploitative intimate imagery and sextortion
- Sexualization of children

[37] https://transparency.fb.com/policies/community-standards/child-sexual-exploitation-abuse-nudity/

- Child nudity
- Non-sexual child abuse

For full policy details, see our Community Standards.

We continuously consult child safety experts to ensure our policies are up-to-date to address the latest safety risks. For example, we updated our child safety policies to clarify that we will remove Facebook profiles, Pages, groups and Instagram accounts that are dedicated to sharing otherwise innocent images of children with captions, hashtags or comments containing inappropriate signs of affection or commentary about the children depicted in the image. We've always removed content that explicitly sexualizes children, but content that isn't explicit and doesn't depict child nudity is harder to define. Under this new policy, while the images alone may not break our rules, the accompanying text can help us better determine whether the content is sexualizing children and if the associated profile, Page, group or account should be removed.

Enforcement

In addition to AI and human content review, as described in our general approach, we have specially trained teams with backgrounds in law enforcement, online safety, analytics, and forensic investigations review potentially violating content and report findings to the National Center for Missing and Exploited Children (**NCMEC**), a not-for-profit organisation whose mission is to help find missing children, reduce child sexual exploitation, and prevent child victimization. NCMEC works with law enforcement agencies around the world to help victims, and support their work in holding offenders accountable. Meta has helped NCMEC develop new software to help prioritise the reports it shares with law enforcement in order to address the most serious cases first.

In addition, we respond to valid legal requests for information from New Zealand law enforcement in accordance with applicable law and our terms of service (see further detail below).

Our policies and enforcement have also been expanded to detect and remove networks that violate our child endangerment policies, similar to our efforts against coordinated inauthentic behaviour and dangerous organisations.

Tools, Products and Resources

We use technology to find child exploitative content and detect possible inappropriate interactions with children or child grooming. Among the detection technologies we use are photo-matching technologies that help us 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. We use these technologies across our public surfaces, as well as on unencrypted information available to us on our private-messaging services, including user reports. We also run these technologies on links from other internet sites shared on our apps and their associated content to detect known child exploitation housed elsewhere on the internet. Not only does this help keep our platforms safer, but it also helps keep the broader internet safer as all violating content is reported to NCMEC.

In addition to photo-matching technology, we're using artificial intelligence and machine learning to proactively detect child nudity and previously unknown and new child-exploitative content, as well as inappropriate interactions with children, sometimes referred to as "grooming."

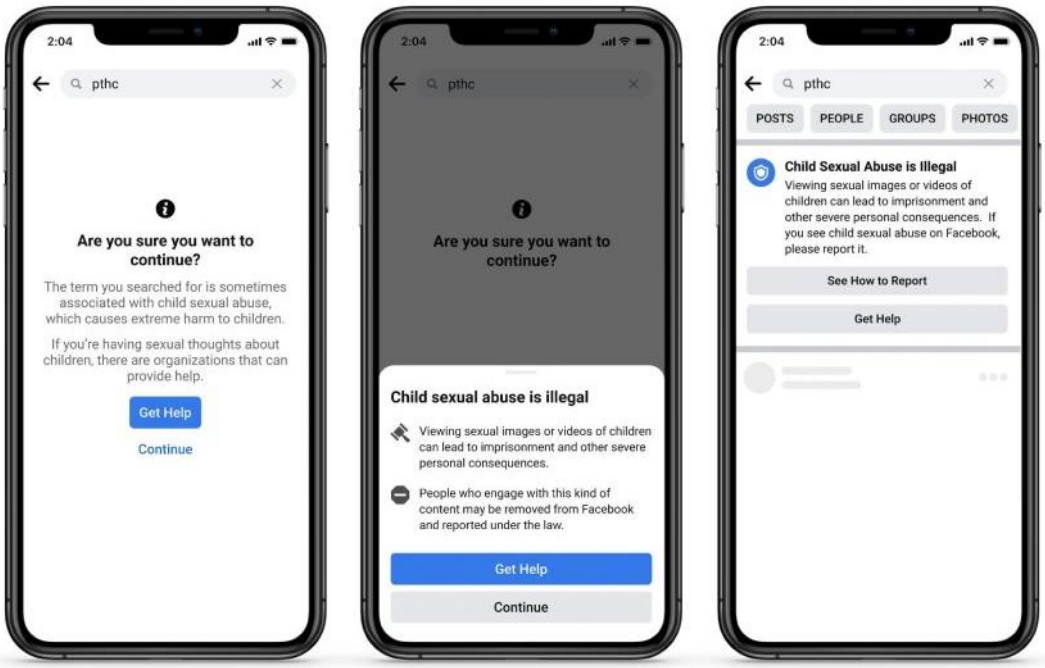
We use these tools along all three prongs of our approach to child protection -- prevention, detection, and response -- and in ways tailored to public spaces like Pages, Groups, and Profiles. The following are additional examples of tools and products that we use to keep children and youth safe from CSEA on our platforms.

- **Detecting violating content.** We have developed two technologies to combat child sexual abuse material (CSAM) — PDQ and TMK+PDQF — to detect identical and near-identical photos and videos. We have made these technologies open source and available to industry partners, small developers and NGOs. The President and CEO of NCMEC John Clarke said, "We're confident that Facebook's generous contribution of this open-source technology will ultimately lead to the identification and rescue of more child sexual abuse victims."[38]

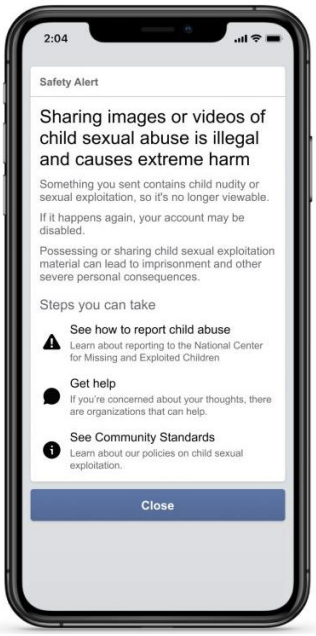
- **Detecting suspicious behaviours.** We've developed new technology that allows us to find accounts that have shown potentially suspicious behaviour and stop those accounts from discovering and interacting with young people's accounts. By "potentially suspicious behaviour", we mean accounts belonging to adults that may, for example, have recently been blocked or reported by a young person. Using this technology, we won't show young people's accounts to these adults who exhibit "potentially suspicious behaviour". If they find young people's accounts by searching for their usernames, they won't be able to follow them. They also won't be able to see comments from young people on other people's posts, nor will they be able to leave comments on young people's posts.

- **Reduce sharing of child exploitative content.** Based on research[39] conducted with NCMEC to improve our understanding of why people may share child exploitation material on our platform, we have introduced new tools and policies targeted at reducing the sharing of this content - one tool is aimed at the potentially malicious searching for child exploitative content and while the other is aimed at the non-malicious sharing of that content. The first is a pop-up shown to people searching for terms on our apps associated with child exploitation and provides ways to get help from offender diversion organisations, while also providing a **warning notice** about the consequences of viewing illegal content.

[38] https://about.fb.com/news/2021/07/instagram-safe-and-private-for-young-people/.
[39] https://research.facebook.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/

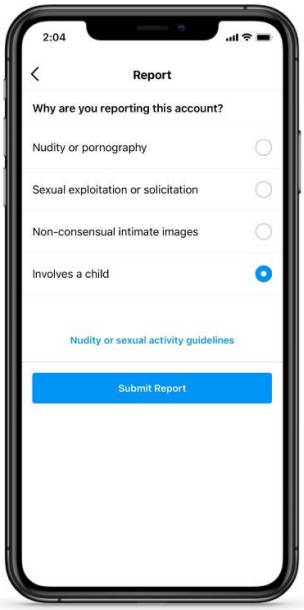


The second is a **safety alert** that informs people who have shared viral, meme child exploitative content about the harm the content causes; warns that it is against our policies; and that there are legal consequences for sharing the material. The safety alert is displayed, and the content is then removed, banked and reported to NCMEC. Accounts that promote this content are also removed. We use insights from the safety alert to help us identify behavioural signals of those who might be at risk of sharing the CSEA content — we then try to educate them on why it is harmful and encourage them not to share it on any surface — public or private.

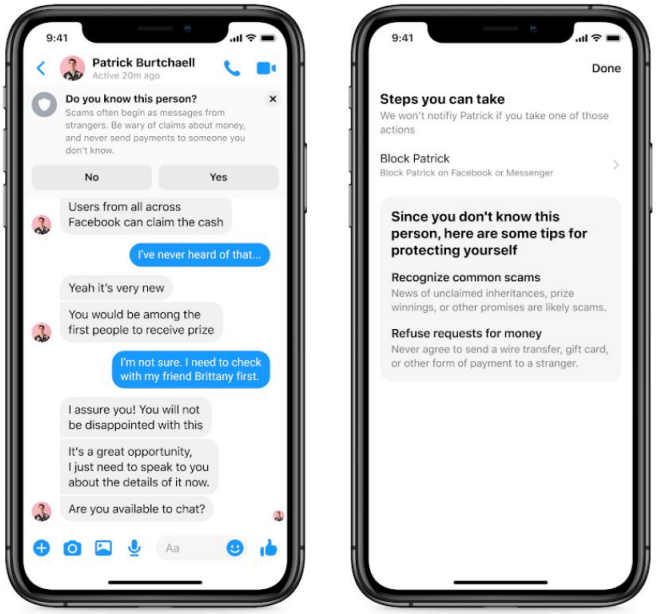


- **User Reporting.** After consultations with child safety experts and organisations, we've made it easier for users to report content that violates our child endangerment policies — we use AI to prioritise and swiftly respond to reports. To

do this, we added the option to choose "involves a child" under the "Nudity & Sexual Activity" category of reporting in more places on Facebook and Instagram. These reports will be prioritised for review. We also started using Google's Content Safety API to help us better prioritise content that may contain child exploitation for our content reviewers to assess.[40]



- **Warning and Safety Notices.** We've introduced warnings and safety notices across our platforms to educate people on who they're engaging with. For example, in Messenger we have introduced safety notices that pop up and provide tips to help people spot suspicious activity or take action to block or ignore someone when something doesn't seem right (see screenshot below). These notices are designed to discourage inappropriate interactions with children and to limit the potential for grooming to occur via Messenger and Instagram.[41]



[40] https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/
[41] https://messengernews.fb.com/2020/05/21/preventing-unwanted-contacts-and-scams-in-messenger/

- **Protecting the Privacy of Youth.** We want to stop young people from hearing from adults they don't know or don't want to hear from, and we believe private accounts are the best way to prevent this from happening. We have longstanding tools that young people can use to protect the privacy of their own accounts, including limiting who can find them, who can send them a friend request and what information is publicly available.
    - Since July 2021, anyone who is under 16 years old in New Zealand is defaulted into a private account when they join Instagram. For young people who already have a public account on Instagram, we show them a notification highlighting the benefits of a private account and explaining how to change their privacy settings.[42]
    - We have designed many of our features to remind minors who they're sharing with and to limit interactions with strangers. This includes protecting sensitive information, such as minors' contact information, school or birthday, from appearing in public searches. Additionally, we take steps to remind minors that they should only accept friend requests from people they know, and we do not allow unconnected adults to message minors.
    - We have invested significantly in AI to detect the age of young users, especially those who may be under 13 and too young to use our apps.[43]
- **Age Assurance.** We require people to be at least 13 years-old to sign up for Facebook or Instagram. Our approach to understanding a user's age aims to strike a balance between protecting people's privacy, wellbeing, and freedom of expression. We take a multi-layered approach to understanding someone's age.
    - We allow anyone to report suspected underage user accounts on Instagram and Facebook. Our content reviewers are trained to flag reported accounts that appear to be underage. If these people are unable to prove they meet our minimum age requirements, we delete their accounts.
    - We require users to provide their date of birth when they register new accounts, a tool called an 'age screen'. Those who enter their age (under 13) are not allowed to sign up. The age screen is age-neutral (i.e. it does not assume that someone is old enough to use our service), and we restrict people who repeatedly try to enter different birthdays into the age screen.
    - Some people may misrepresent their age online and enforcement can be challenging in this space. Our technology allows us to estimate people's ages, using multiple signals, e.g. we look at things like people wishing a user happy birthday and the age written in those posts. We also look at the age users have shared across apps, e.g. if a user has shared their birthday on Facebook, we'll use the same for linked accounts on Instagram.

[42] https://about.fb.com/news/2021/07/instagram-safe-and-private-for-young-people/
[43] https://about.fb.com/news/2021/07/age-verification/

    - Between July and September 2021, Meta removed more than 2.6 million accounts on Facebook and 850,000 accounts on Instagram globally because they were unable to meet our minimum age requirement.[44]

- **Age-Appropriate Restrictions.** For those users that we know or suspect are between the ages of 13 and 18, we take a number of steps to ensure they have an age-appropriate experience on Facebook and Instagram:

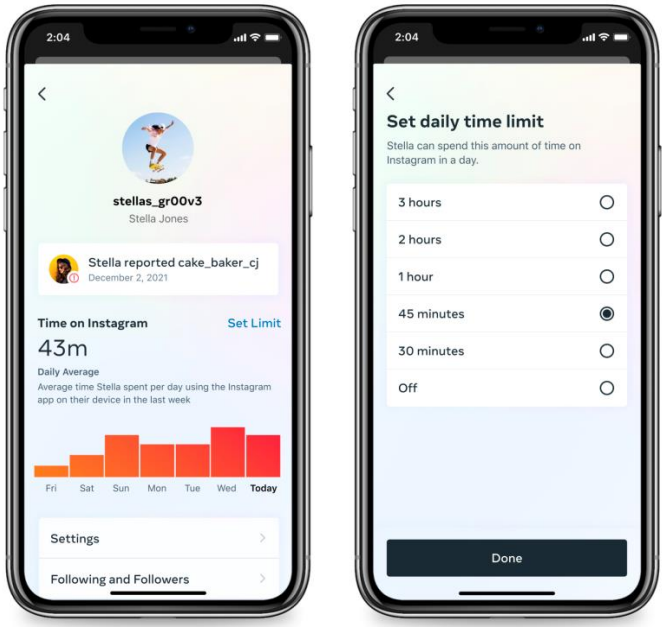
    - We place a range of default limits on a minor's accounts. For example, minor profiles cannot be found on Facebook, Instagram, or search engines off our platform; Post and Story audiences are defaulted to Friends (rather than public); and Location is defaulted off.

    - We only allow advertisers targeting ads to people under 18 (or older in certain countries), based on their age, gender and location. This means certain targeting options, like those based on minors' interests or their activity on other apps and websites, are not available to advertisers. This is in addition to age-gating controls made available for those advertisers who publish age-sensitive ads or content (such as related to gambling).

- **Parental Supervision.** In addition to the responsibility of industry to invest in safety, parents and carers play a vital role in ensuring the safety of young people online. We work to provide tools and resources for parents and guardians so they can guide and support their children and teens. In 2020, we launched Messenger Kids in New Zealand which places parental controls at the heart of the experience, so that younger users can connect with their friends, while parents can monitor their children's privacy and security controls. Parents manage who their child interacts with and can monitor their child's activity in the app through the Parent Dashboard, where they can also download their child's information at any time. Parents and guardians may also view how much time their teens spend on Instagram and set time limits.[45] Teens also have the option to notify their parents if they report someone, giving their parents the opportunity to talk about it with them.

[44] A Mosseri, Hearing Before the United States Senate Committee on Commerce, Science, and Transportation Subcommittee on Consumer Protection, Product Safety, and Data Security, 8 December 2021, https://www.commerce.senate.gov/services/files/3FC55DF6-102F-4571-B6B4-01D2D2C6F0D0

[45] https://about.fb.com/news/2021/12/new-teen-safety-tools-on-instagram/



- **Resources for Parents and Youths.** We offer a number of informative resources to parents, children, and educators to help increase awareness and online safety education. Specifically:
    - Instagram Safety and Wellbeing Hub[46] and the Facebook Safety Center[47] which provide users with tools to stay safe, secure their accounts and protect their information.
    - Youth Portal provides a central place for teens to access education on our tools and products, first person accounts from teens about how they're using technologies, tips on security and reporting, and advice on how to use social media safely.[48]
    - Get Digital Hub, a digital citizenship and wellbeing program which provides schools and families with lesson plans and activities to help build the core competencies and skills young people need to navigate the digital world in safe ways.[49]

Partnerships

Child protection requires a global and comprehensive response from industry, law enforcement, government, civil society, and families — we have and continue to work with all these stakeholders to strengthen the child safety ecosystem. We collaborate across industry through organisations like the Tech Coalition, an alliance of global tech companies who are working together to combat child sexual exploitation and abuse online.

- In New Zealand, we partner with Netsafe, across a range of areas, including supporting Netsafety Week and developing online safety resources. In 2019, we partnered with Netsafe to create a localised Instagram Safety Guide for parents

[46] https://about.instagram.com/safety
[47] https://www.facebook.com/safety/tools
[48] https://www.facebook.com/safety/youth/
[49] https://www.facebook.com/fbgetdigital

(which will be updated in 2023).

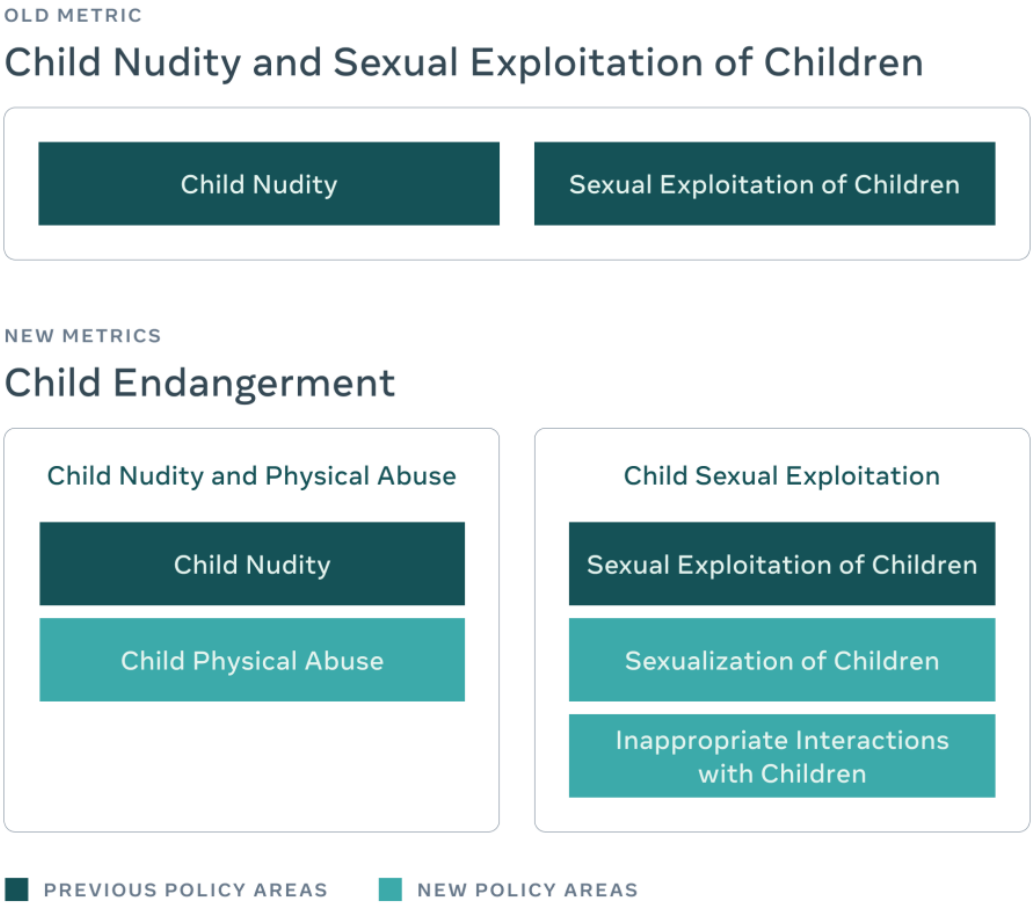
- We cooperate with New Zealand law enforcement, including New Zealand Police and the Department of Internal Affairs, to combat abuse and prevent real-world harm. We provide a dedicated law enforcement program and respond to valid legal requests for information from New Zealand law enforcement, including with respect to investigations into CSEA,, in accordance with applicable law and our terms of service. We also provide an emergency response channel for law enforcement to seek information in circumstances where there is a risk of death or imminent bodily harm. We engage with New Zealand law enforcement regarding the cases we refer to NCMEC, and have incorporated feedback provided by New Zealand authorities to ensure the reports contain the information that is most useful to support investigations.
- In Q4 2021, we received 127 legal process requests and 132 emergency disclosure requests from New Zealand law enforcement, in relation to 417 users/accounts.
- We work closely with our Safety Advisory Board, which is comprised of 11 leading online safety nonprofits (including NetSafe), and with over 400 safety experts and NGOs from around the world, including specialists in combating child-sexual exploitation and aiding its victims. Our efforts include developing industry best practices, building and sharing technology to fight online child exploitation, and supporting victim services, among other things.
- We launched AMBER Alerts on Facebook in 2015 to help families and authorities successfully recover missing children and have since expanded the program to over 20 countries, including in New Zealand. People in a designated search area where local law enforcement has activated an AMBER Alert will see the alert in their News Feed. The alert includes a photo of the missing child, a description, the location of the abduction, and any other pertinent, available information. People can share the alert with friends to spread awareness. We know the chances of finding a missing child increase when more people are on the lookout, especially in the critical first hours. Our goal is to help get these alerts out quickly to the people who are in the best position to help.
- In 2020, Meta joined Google and other member companies of the Tech Coalition to launch Project Protect: A plan to combat online child sexual abuse. This includes a renewed commitment and investment from the Tech Coalition and expands its scope and impact to protecting kids online. Project Protect focuses on five key areas: tech innovation, collective action, independent research, information and knowledge sharing, and transparency and accountability.

Transparency & Accountability

We publish a quarterly Community Standards Enforcement Report that discloses metrics on the effectiveness of our policies and processes in reducing the prevalence of content that endangers children, such as content that contains nudity or physical abuse or content that sexually exploits children on Facebook and Instagram.

In the second quarter of 2021, we expanded the metric for child endangerment — which originally only reported on child nudity and sexual exploitation — to include 1) nudity and physical abuse and 2) sexual exploitation. The global figure below represents child nudity

and sexual exploitation only in the first quarter, and the expanded metric for Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation in the remaining three quarters of 2021.



The table below shows the pieces of child endangerment content that we took action on globally in 2021 and the proactive rate of content we detected before people reported it.[50]

Period	Child Nudity and Sexual Exploitation	Child Nudity and Physical Abuse	Child Sexual Exploitation
Jan-Mar	Facebook: 5 million with proactive rate over 98% Instagram: 812,000 with proactive rate over 98%	not available	not available
Apr-Jun	changed approach	Facebook: 2.3 million with proactive rate over 97% Instagram: 458,000 with proactive rate over 95%	Facebook: 25.6 million with proactive rate over 99% Instagram: 1.4 million with proactive rate over 96%
Jul-Sep	changed approach	Facebook: 1.8 million with proactive rate over 97% Instagram: 527,000 with	Facebook: 21.2 million with proactive rate over 99% Instagram: 1.6 million with

[50] https://transparency.fb.com/data/community-standards-enforcement/child-nudity-and-sexual-exploitation/facebook/

Period	Child Nudity and Sexual Exploitation	Child Nudity and Physical Abuse	Child Sexual Exploitation
		proactive rate over 92%	proactive rate over 96%
Oct-Dec	changed approach	Facebook: 1.8 million with proactive rate over 97% Instagram: 983,000 with proactive rate over 95%	Facebook: 19.8 million with proactive rate over 99% Instagram: 2.6 million with proactive rate over 97%

For New Zealand, from April to December 2021:

- We took action on over 14 thousand 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 5 thousand pieces of content on Instagram in New Zealand for violating our Child Sexual Exploitation policies. Over 87% of this content was detected proactively before people reported it to us.

Note: Due to the changes in our approach to the metrics for child endangerment content, the metrics below only reflect the period April to December (Q2-Q4) 2021.

Cyberbullying or Harassment

When it comes to bullying and harassment, context matters. It can be hard to tell the difference between a bullying comment and a light-hearted jest without knowing the people involved or the nuance of the situation. This is why we continuously explore new ways to tackle the issue.

We are building new tools, updating our policies, and investing in detection technology to ensure we are proactively tackling the problem as best we can, as we know how important it is to get this right. This includes having expert teams who review reports of bullying and harassment and AI technology to detect and take action on violating content.

Where we do not have language support, we make use of translation services but also compile lists of slur words that can trigger review. We rely on reports from our community and our trusted partners around the world; in New Zealand our trusted partner is NetSafe, though we also receive reports from other NGOs and public entities.

Policies

Bullying and harassment happen in many places and come in many different forms, from making threats and 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 our platforms.

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 we remove content that's meant to degrade or shame, including, for example, claims about someone's personal 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.

We continue to update our policy to adjust for the gendered and culturally specific nature that some forms of online harassment and abuse can occur, especially for women. In July 2019, for example, we expanded our bullying and harassment policy to enforce more strictly on cursing that uses female-gendered terms. We are looking forward to exploring these areas further, via the mechanisms in the Code, concerning the online abuse of Māori in New Zealand.

Our policies have also been updated to provide more protections for female public figures, to combat degrading or sexualised attacks. In 2021, we announced further changes to remove unwanted sexualised commentary and repeated content that is sexually harassing.[51] Because what is "unwanted" can be subjective, we rely on additional context from the individual experiencing the abuse to take action. We made these changes because attacks like these can weaponise a public figure's appearance, which is unnecessary and often not related to the work these public figures represent.

The full details of our bullying and harassment policy can be found here.[52]

Enforcement

It is important to note that bullying and harassment is often contextual and personal, making enforcement at scale challenging when compared to some other harm categories. In certain instances, we require self-reporting because it helps us understand that the person targeted feels bullied or harassed. And, context and intent matter, especially if someone has shared something in order to condemn or raise awareness.

We use human review and developed AI systems to identify and take action on many types of bullying and harassment across our platforms. This includes removing posts, accounts, Pages, Groups and events for violating our Community Standards or Guidelines. However, as mentioned, given the highly contextual nature, using technology to proactively detect these behaviours can be more challenging than other types of violations. That's why we rely on people to report this behaviour to us so we can identify and remove it.

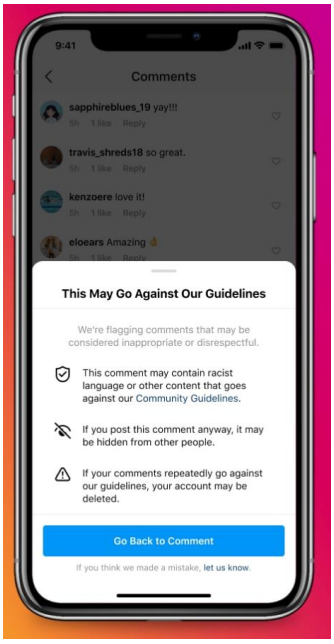
Additionally, bullying and harassment can cut across different types of abuse; for example, a racial slur could be used to bully or harass an individual, which we would remove for violating our hate speech policy. Other policies that can be related to bullying and harassment include hate speech and violence and incitement.

Tools, Products and Resources

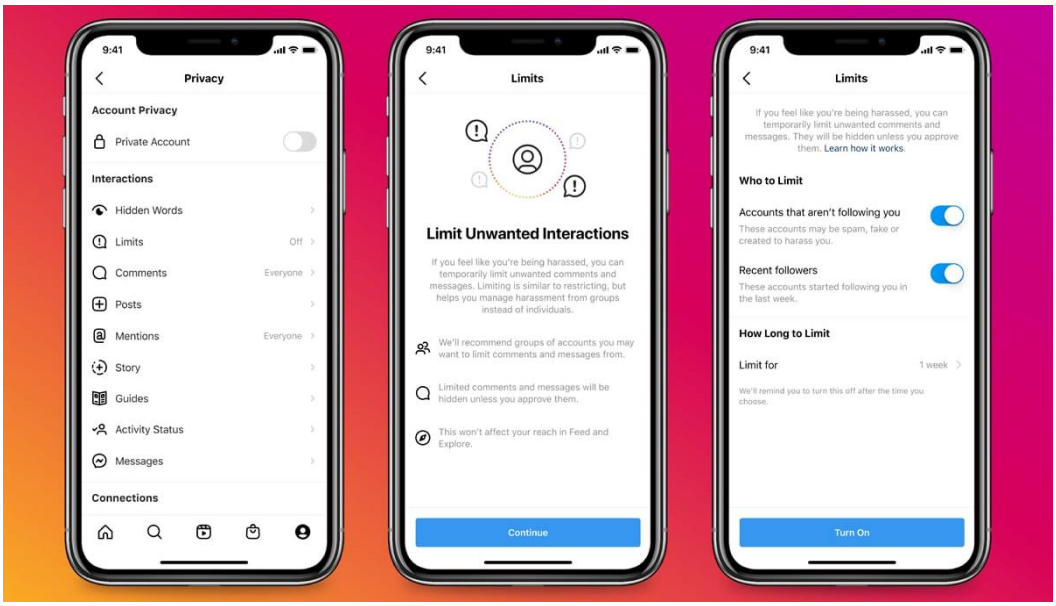
Even if content does not violate our Community Standards, people may prefer to not see it. They may also want to take steps in order to control their individual experience on our platform. The following are some examples of products and tools implemented empower users to protect themselves from bullying and harassment:

[51] https://about.fb.com/news/2021/10/advancing-online-bullying-harassment-policies/
[52] https://transparency.fb.com/en-gb/policies/community-standards/bullying-harassment/

- **Warnings to Discourage Harassment**. We deployed a tool on Facebook and Instagram that sends a warning to educate and discourage people from posting or commenting in ways that could be bullying and harassment. We found that after viewing these warnings on Instagram, about 50 percent of the time the comment was edited or deleted by the user.[53]



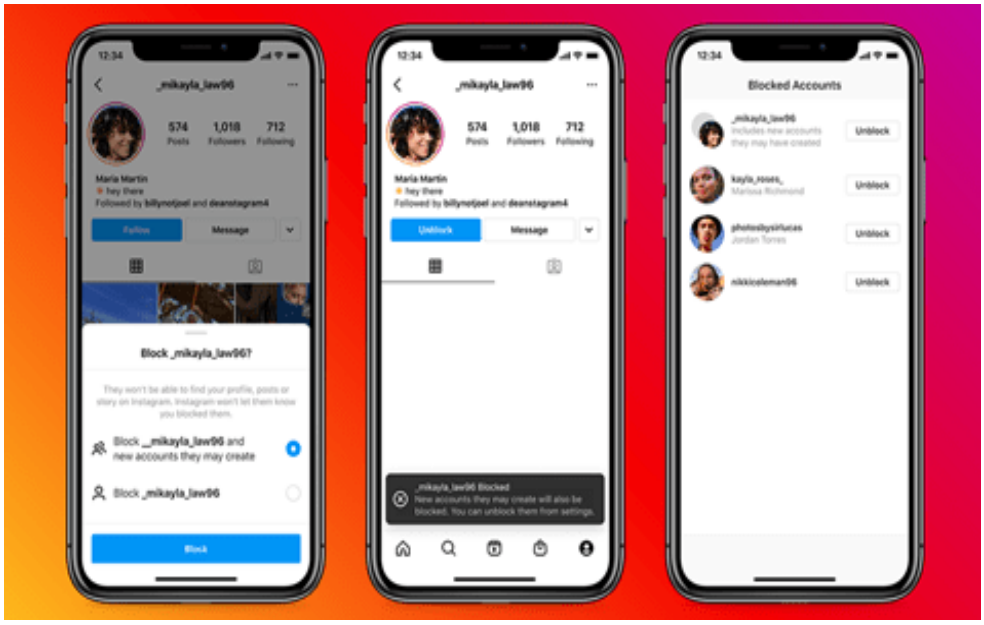
- **Limit Comments.** The Limits tool on Instagram allows users to automatically hide comments and Direct Messaging requests from people who don't follow them, or who only recently followed them, to help manage an unexpected influx of unwanted contact.[54] We developed and launched this tool in partnership with sports organisations , to help protect players from racist abuse. This feature is available globally, including in New Zealand.



[53] https://about.fb.com/news/2021/11/how-meta-addresses-bullying-harassment/
[54] https://about.instagram.com/blog/announcements/introducing-new-ways-to-protect-our-community-from-abuse

- **Restrict Commenting Audience.** Users can control their commenting audience for a public post by choosing from a menu of options. By adjusting the commenting audience, users can further control how they want to invite conversation onto their public posts, and limit potentially unwanted interactions.[55]

- **Filter Abusive Messages.** Because direct messages on Instagram are private conversations, we don't proactively look for bullying and harassment the same way we do on public surfaces. That's why we introduced a new tool which, when turned on, will automatically filter message requests containing offensive words, phrases and emojis, so users never have to see them.[56]

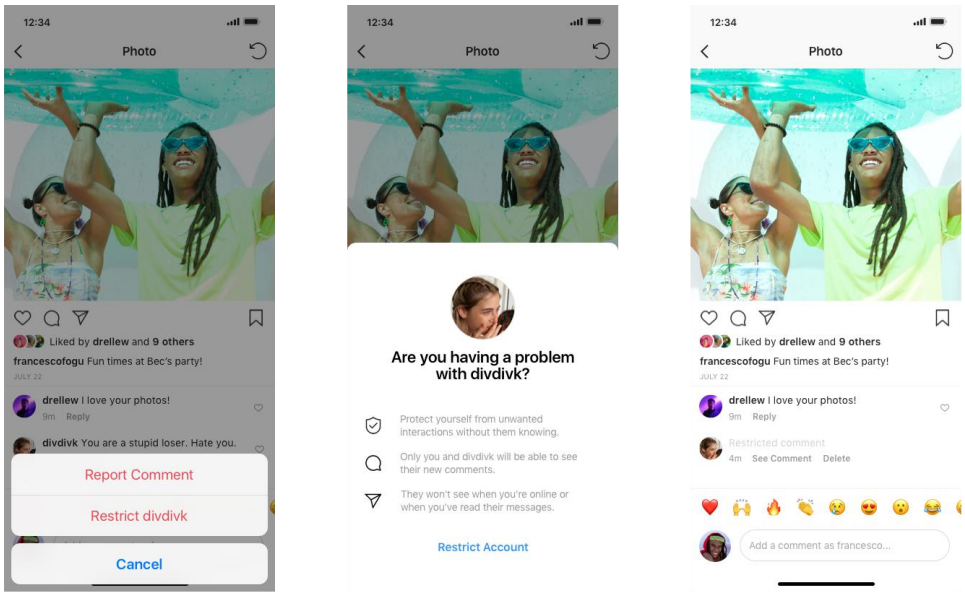


- **Restrict What People Can See About You.** We've created a Restrict tool in Instagram[57] where comments on people's posts from a person they have restricted will only be visible to that person. Direct messages will automatically move to a separate Message Requests folder, and people will not receive notifications from a restricted account. People can still view the messages but the restricted account will not be able to see when their direct messages were read or when someone is active on Instagram.

[55] https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/
[56] Ibid.
[57] https://about.instagram.com/blog/announcements/stand-up-against-bullying-with-restrict

- **Resources on Women, Parents and Youths.** We offer a number of resources to educate users on how to protect themselves from bullying and harassment. Specifically:

    - Instagram Safety and Wellbeing Hub[58] and the Facebook Safety Center[59] provide users with tools to stay safe, secure their accounts and protect their information.



    - Bullying Prevention Hub is a resource for teenagers, 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.

    - Family Center provides parents with tools and resources to help support their teens' online experience. Parents may oversee their teens' accounts within Meta technologies, set up and use supervision tools, and access resources on how to communicate with their teens about internet use. The Family Center includes an education hub where parents and guardians can access resources from experts and review helpful articles, videos and tips on topics like how to talk to teens about social media.[60]

    - Women's Safety Hub is a dedicated safety page for women that offers information about tools and resources that can help women feel safe online.[61]

Partnerships

Partnerships are especially important in raising awareness and educating users on bullying and harassment issues and to know how to report this type of activity.

- We work with Women's Refuge New Zealand to promote their messaging campaigns around family violence, especially during COVID-19 lockdowns, so they

[58] https://about.instagram.com/safety
[59] https://www.facebook.com/safety/tools
[60] https://about.fb.com/news/2022/03/parental-supervision-tools-instagram-vr/
[61] https://www.facebook.com/safety/womenssafety

can reach vulnerable New Zealanders who may need their help. We also supported the "She is not your rehab" haka against violence campaign that rolled out during the 2020 lockdown when domestic violence numbers were on the rise.

- In 2021 we announced our Global Women's Safety Expert Advisors,[62] a group of 12 nonprofit leaders, activists and academic experts to help us develop new policies, products and programs that better support the women who use our apps. This expert group includes Dr Asher Flynn, an Associate Professor of Criminology at Monash University and the Vice President of the Australian and New Zealand Society of Criminology. Dr Flynn's work focusses on AI-facilitated abuse, deepfakes, gendered violence and image-based sexual abuse.

- We have invested in industry-leading initiatives to combat the non-consensual sharing of intimate images (NCII), often referred to as 'revenge porn'. It has long been our policy on Facebook and Instagram to remove NCII. In 2021 we launched an initiative in New Zealand, partnering with Netsafe, to help victims proactively stop the proliferation of their intimate images. [63] Through the use of hash technology and partner support, a user can proactively share an image to stop it from being uploaded by other parties for nefarious purposes.[64] Following the success of this pilot, we recently launched the expansion of the program globally, known as StopNCII.org. StopNCII.org operates in partnership with more than 50 non-governmental organisations around the world, including NetSafe.

- We supported the recent New Zealand legislative change instigated by former MP Louisa Wall's private member's bill on specifically criminalising NCII.[65]

- We also work with third party experts to develop resources specifically designed to promote women's safety. In partnership with global NGO Thorn, we launched the Stop Sextortion Hub, which provides resources for teens, caregivers and educators seeking support and information related to sextortion.

Transparency & Accountability

We publish a quarterly Community Standards Enforcement Report that discloses metrics on the effectiveness of our policies and processes in reducing the prevalence of bullying and harassment content on Facebook and Instagram. The table below shows the pieces of content that we took action on globally in 2021 and the proactive rate of content detected before people reported it.[66]

[62] https://about.fb.com/news/2021/06/partnering-with-experts-to-promote-womens-safety/
[63] https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images/
[64] https://www.facebook.com/safety/notwithoutmyconsent/pilot/how-it-works
[65] *Facebook submission on the Harmful Digital Communications (Unauthorised Posting of Intimate Visual Recording) Amendment Bill,* https://www.parliament.nz/resource/en-NZ/53SCJU_EVI_99360_JU1444/abed582a4374178a212be579736ed039a19aaa6f.
[66] https://transparency.fb.com/data/community-standards-enforcement/bullying-and-harassment/facebook/

Period	Facebook	Instagram
Jan-Mar	8.8 million with proactive rate over 54%	5.6 million with proactive rate over 78%
Apr-Jun	7.9 million with proactive rate over 54%	4.5 million with proactive rate over 71%
Jul-Sep	9.2 million with proactive rate over 59%	7.8 million with proactive rate over 83%
Oct-Dec	8.2 million with proactive rate over 58%	6.6 million with proactive rate over 82%

For New Zealand, in 2021:

- We took action on over 140 thousand pieces of content on Facebook in New Zealand for violating our Bullying & Harassment policy. Over 78% of this content was detected proactively before people reported it to us.
- We took action on over 130 thousand pieces of content on Instagram in New Zealand for violating our Bullying & Harassment policy. Over 86% of this content was detected proactively before people reported it to us.

Hate Speech

We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That is why 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.

Our policies for hate speech provide a greater degree of specificity, ability to adapt more quickly, and in some countries, including in New Zealand, broader protections than are available under local legal settings.

The problem of hate speech, however, is not just for platforms to solve. It is a whole-of-society problem in which governments have a role to play, in partnership with experts, industry and the broader community. Learn more about Meta's approach to hate speech here.

Policies

We define hate speech as a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.

We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanising comparisons that have historically been used to attack, intimidate or exclude specific groups, and that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic.

We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we

provide some protections for characteristics such as occupation, when they're referenced along with a protected characteristic. Sometimes, based on local nuance, we consider certain words or phrases as frequently used proxies for PC groups.

We also prohibit the usage of slurs that are used to attack people on the basis of their protected characteristics. However, we recognise that people sometimes share content that includes slurs or someone else's hate speech to condemn it or raise awareness. In other cases, speech, including slurs, that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If the intention is unclear, we may remove content.

However, we recognise that people sometimes share content that includes slurs or someone else's hate speech to condemn it or raise awareness. In other cases, speech, including slurs, that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If the intention is unclear, we may remove content.

We have made a number of changes to expand our hate speech policy in our Community Standards. These include:

- the development of a new hateful stereotypes policy, which will in the first instance prohibit content depicting blackface and stereotypes that Jewish people run the world;[67]
- expansions in our ads policies to better protect immigrants, migrants, refugees and asylum seekers from hateful claims;[68]
- expansions in our ads policies to prohibit claims that a group is a threat to the safety, health or survival of others on the basis of that group's race, ethnicity, national origin, religious affiliation, sexual orientation, gender, gender identity, serious disease or disability; [69]
- amendments to our policy to remove any claims that deny or distort the Holocaust, on the basis of expert consultation and research.[70]

The full details of our hate speech policy can be found here.[71]

Enforcement

Enforcing our policies against hate speech has historically been the most challenging content for artificial intelligence to detect because it is dependent on nuance, history, language, religion and changing cultural norms.

We need to be confident that something is hate speech before we remove it. If something might be hate speech but we're not confident enough that it meets the bar for removal, our technology may reduce the content's distribution or won't recommend Groups, Pages or

[67] https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/
[68] https://about.fb.com/news/2020/06/meeting-unique-elections-challenges/
[69] Ibid.
[70] https://about.fb.com/news/2020/10/removing-holocaust-denial-content/
[71] https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

people that regularly post content that is likely to violate our policies. We also use technology to flag content for further review. As such, we have a high threshold for automatically removing content.[72]

In 2016, the vast majority of our content removals were based on what users reported to us. Our investment in AI is evident from the increasing percentage of hate speech content we have been detecting proactively. At the end of 2017, for example, less than 25 per cent of hate speech content we removed on Facebook were detected proactively. This figure has progressively increased over time; at the end of 2021, over 95% of hate content on Facebook have been proactively detected.

This improvement in our detection ability was accompanied by a stark increase in the total volume of hate speech content we have removed. At the end of 2017, we removed 1.6 million pieces of hate speech content on Facebook; by the end of 2021, we removed 17.4 million pieces of content. Information on our progress on AI and hate speech detection can be found here.[73]

Aside from our Repeat Violators policy that allows us to take action on pages, groups, profiles, and accounts that repeatedly violate our hate speech policy, we have been working with teams across the company to expand our network disruption efforts so we can address threats that come from groups of authentic accounts coordinating on our platform to cause social harm. For example, we removed a network of Facebook and Instagram accounts, Pages and Groups associated with the Querdenken movement (which is linked to off-platform violence and other social harms) for engaging in *coordinated efforts* to repeatedly violate our Community Standards, including posting hate speech and incitement to violence. We also blocked their domains from being shared on our platform.[74]

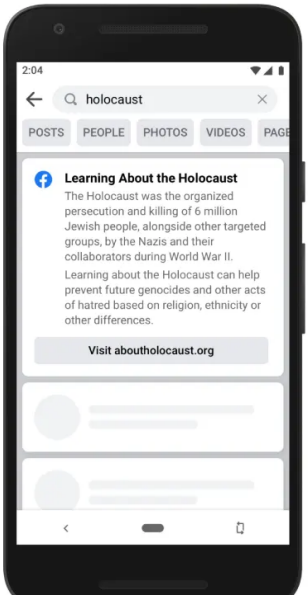
How We Reduce the Prevalence of Hate Speech

Our efforts are having a big impact on reducing how much hate speech people see on Facebook, with the prevalence of hate speech currently at about 0.05% of content viewed, or about 5 views per every 10,000, down by almost 50% in the last three quarters. This chart is illustrative of the efforts taken to combat hate speech.

[72] https://about.fb.com/news/2021/10/hate-speech-prevalence-dropped-facebook/
[73] https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/
[74] https://about.fb.com/news/2021/09/removing-new-types-of-harmful-networks/

Tools, Products and Resources

Many of our tools, products and resources are aimed at addressing a wide range of online safety and harms issues, including hate speech. Tools have been made available such as adding warning screens and checks to proactively prevent users from posting content that may be hateful. We also deploy counterspeech programs and resources to help counter hate speech. For example, given the well-documented rise in anti-Semitism globally, we've taken steps to help educate people about the events that led to the Holocaust and the genocide of one-third of the Jewish people, by connecting people with credible information about the Holocaust.



We're taking these steps given the well-documented rise in anti-Semitism globally and the alarming level of ignorance about the Holocaust, especially among young people. We want to help peple learn about the events that led to the Holocaust and the genocide of one-third of the Jewish people.

Partnerships

Aside from the many partnerships we have around the world, we supported the New Zealand Human Rights Commission on the first and second version of their 'Dial it Down' Campaign, which has a specific focus on encouraging people to be respectful and positive online. We also work with the Sakinah Community Trust, supporting their efforts to promote community cohesion through Unity Week.

Transparency & Accountability

We publish a quarterly Community Standards Enforcement Report that discloses metrics on the effectiveness of our policies and processes in reducing the prevalence of hate speech content on Facebook and Instagram. The table below shows the pieces of content that we took action on globally in 2021 and the proactive rate of content detected before people reported it.[75]

[75] https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/

Period	Facebook	Instagram
Jan-Mar	25.2 million with proactive rate over 96%	6.3 million with proactive rate over 93%
Apr-Jun	31.5 million with proactive rate over 97%	9.8 million with proactive rate over 95%
Jul-Sep	22.3 million with proactive rate over 96%	6 million with proactive rate over 93%
Oct-Dec	17.4 million with proactive rate over 95%	3.8 million with proactive rate over 91%

For New Zealand, in 2021:

- We took action on over 61 thousand pieces of content on Facebook in New Zealand for violating our Hate Speech policy. Over 90% of this content was detected proactively before people reported it to us.
- We took action on over 74 thousand 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.

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 challenge is made worse in places where there is a heightened risk of conflict and 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.

Policies

We aim to prevent potential offline harm that may be related to content on our platforms. 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. Our Violence and Incitement[76] policy allows us to 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.

Under our Dangerous Individuals and Organisations[77] policy that prohibits the sharing of content that glorifies violence and terrorist content. In an effort to prevent and disrupt real-world harm, we do not allow organisations or individuals that proclaim a violent mission or are engaged in violence to have a presence on our platforms. We assess these entities

[76] https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/
[77] https://transparency.fb.com/en-gb/policies/community-standards/dangerous-individuals-organizations/

based on their behaviour both online and offline, most significantly, their ties to violence. Under this policy, we designate individuals, organisations, and networks of people. In 2019, we strengthened the policy by banning praise, support and representation of white nationalism and separatism on Facebook and Instagram.[78] We recognize that users may share content that includes references to designated dangerous organisations and individuals to report on, condemn, or neutrally discuss them or their activities. Our policies are designed to allow room for these types of discussions while simultaneously limiting risks of potential offline harm. We thus require people to clearly indicate their intent when creating or sharing such content. If a user's intention is ambiguous or unclear, we default to removing content. And, in line with international human rights law, our policies allow discussions about the human rights of designated individuals or members of designated dangerous entities, unless the content includes other praise, substantive support, or representation of designated entities or other policy violations, such as incitement to violence.

We are continuously developing and evaluating our policies to prohibit harmful content and behaviour: refining our policies to address evolving nuances of hate speech, identifying groups at heightened risk of violence or perpetrators of atrocities and human rights abusers.[79]

Enforcement

We have dedicated teams spanning product, engineering, policy, research and operations to better understand and address the way social media is used to incite violence, especially in countries experiencing conflict. Many of these individuals have experience working on conflict, human rights and humanitarian issues, as well as addressing areas like misinformation, hate speech and polarisation.

There is a range of content that we might remove under our violence and incitement policy where someone may advocate for violence or has made a statement of intent to commit violence. Due to the potentially harmful nature of content attempting to incite violence, we over-index on safety and remove such content even if it is unclear whether the content is in jest. This could range from something serious such as instructions on how to use weapons to cause injury to a joke where one friend says to another “I’ll kill you!”. In instances where necessary, we also work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety.

When it comes to dangerous organisations and individuals, designations are divided into three tiers that indicate the level of enforcement, with Tier 1 resulting in the most extensive enforcement because we believe these entities have the most direct ties to offline harm. More information on the tiers under our dangerous organisations and individuals policy can be found here.[80] More than 250 white supremacist organisations have been banned from our platforms under this policy, and we use a combination of AI and human expertise to remove content praising or supporting these organisations.[81]

In New Zealand, we designated the white nationalist group - Action Zealandia - as a dangerous organisation which has the effect of banning the group, or successor groups

[78] https://newsroom.fb.com/news/2019/03/standing-against-hate/
[79] https://about.fb.com/news/2021/02/an-update-on-myanmar/
[80] https://transparency.fb.com/en-gb/policies/community-standards/dangerous-individuals-organizations/
[81] https://about.fb.com/news/2020/10/removing-holocaust-denial-content/

from having a presence on any of our services. This adds to previous people and Groups in the trans-Tasman region we've removed including Neil Erikson, Tom Sewell, the Lads Society, the United Patriots Front, True Blue Crew and the Antipodean Resistance from Facebook and Instagram for violating our policies. These Pages, Groups and individual users posted material that uses dehumanising language towards groups of people, and in some instances, calls for organised hate. These actions occurred prior to any government designation.

Tools, Products and Resources

When we assess a situation to have a high risk of violence, we also have the ability to activate other tools such as:

- Temporarily designating a place as a high-risk location which allows us to take further action on calls for violence.
- Increasing the requirement of Group administrators to review and approve posts before they can be posted by members
- Working with law enforcement on emergency disclosures of information in circumstances where there is a risk of death or imminent bodily harm.

We recently deployed a number of these measures, among others, when responding to the Wellington Parliamentary occupation and riots in early 2022.

Partnerships

Understanding and engaging with local contexts and communities is imperative in addressing and reducing the prevalence of harmful content that incites violence. Over the past few years, we've expanded these relationships with law enforcement and local civil society organisations.

In New Zealand, we cooperate with law enforcement, including New Zealand Police and the Department of Internal Affairs, to respond to online content that incites violence and prevents real-world harm. We provide a dedicated law enforcement program and respond to valid legal requests from New Zealand law enforcement for information in accordance with applicable law and our terms of service. In addition, we provide an emergency response channel for law enforcement to seek information in circumstances where there is a risk of death or imminent bodily harm.

We also support the New Zealand Police Tech Coordinators program through regular training and engagement, ensuring police across the country have necessary expertise to respond to online harms. There are instances where we have cooperated with law enforcement on a range of issues concerning risks of death or imminent bodily injury.

Transparency & Accountability

We publish a quarterly Community Standards Enforcement Report that discloses metrics on the effectiveness of our policies and processes in reducing the prevalence of content that incites violence on Facebook and Instagram. The table below shows the pieces of

content that we took action on globally in 2021 and the proactive rate of content we detected before people reported it.[82]

Period	Facebook	Instagram
Jan-Mar	not available	not available
Apr-Jun	not available	not available
Jul-Sep	13.6 million with proactive rate over 96%	3.3 million with proactive rate over 96%
Oct-Dec	12.4 million with proactive rate over 96%	2.6 million with proactive rate over 96%

For New Zealand, from July to December 2021:

- We took action on over 51 thousand pieces of content on Facebook in New Zealand for violating our Violence & Incitement policy. Over 90% of this content was detected proactively before people reported it to us.
- We took action over 31 thousand pieces of content on Instagram in New Zealand for violating our Violence & Incitement policy. Over 97% of this content was detected proactively before people reported it to us.

Note: We added these metrics for our Violence and Incitement policy in Q3 2021. The global and country metrics above represent the period of July to December 2021.

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.

Policies

To protect users from disturbing imagery, 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. This may include:

- Violent or graphic content that meets our list of exceptions (for example, it provides evidence of human rights abuses or an act of terrorism).

[82] https://transparency.fb.com/data/community-standards-enforcement/violence-incitement/facebook/

- Adult sexual activity or nudity that meets our list of exceptions (for example, culturally significant fictional videos that depict non-consensual sexual touching).
- Suicide or self-injury content that is deemed to be newsworthy.
- Imagery of non-sexual child abuse, where law enforcement or child protection stakeholders ask us to keep the video visible for the purposes of finding the child.

An example of this policy working was in 2021, following the Lynn Mall supermarket terrorist attack, where we restricted (and in some cases removed) graphic and potentially privacy-violating content, while allowing some bystander footage to be viewed. We engaged with the New Zealand Police and the Chief Censor's Office to ensure we were doing our part during this crisis to keep our users safe.

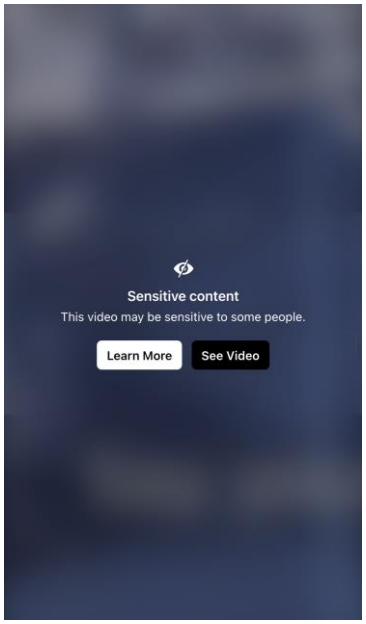
Enforcement

We continue to improve our AI to detect and take action on posts that are likely to contain violent or graphic content. We also improved the context we provide to human reviewers so that they can make informed decisions and we have built systems to help us contact first responders to get help on the ground.

We remove content that glorifies violence or celebrates the suffering or humiliation of others on Facebook and Instagram. We do allow people to share some graphic content to raise awareness about current events and issues. In these cases, we may hide the content from people under 18 and cover it with a warning for those over 18, so people are aware it is graphic or violent before they choose to view it.

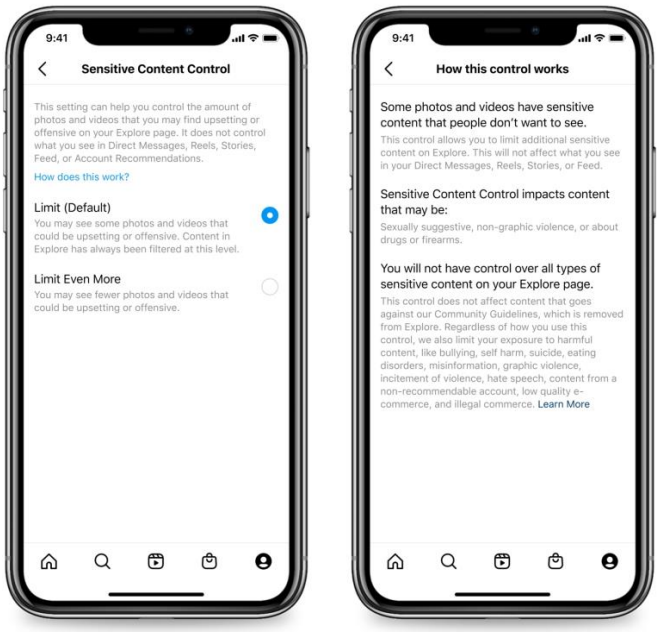
There are different levels of actions that we take for violent and graphic content. We may remove the content; apply a warning screen to alert people that the content is disturbing (in which only users 18 and older may view); or apply a label so that people are aware the content may be sensitive.

For example, once a piece of content is identified as 'disturbing' or 'sensitive' we apply a warning label that limits users from seeing the content unless they click through, shown below. The content will not appear, or present the option of viewing it, for users who are under the age of 18.



Tools, Products and Resources

We have built a range of tools that allow people to control what they see on our platforms as well as warning notices, screens and labels, that have been discussed throughout this report. Specific to graphic and violent imagery, we recognise that people have different sensitivities, so we introduced in 2021 a tool that allows users to decide how much sensitive content shows up in Explore on Instagram. The Sensitive Content Control has three options: Allow, Limit and Limit Even More. "Limit" is the default state for everyone and based on our Recommendation Guidelines, "Allow" enables people to see more sensitive content, whereas "Limit Even More" means they see less of this content than the default state. The "Allow" option is unavailable to people under the age of 18.



The Sensitive Content Control focuses on content that potentially makes people feel unsafe, such as:

- Content that may depict violence, such as people fighting. (Note: We remove graphically violent content.)
- Content that may be sexually explicit or suggestive, such as pictures of people in see-through clothing. (Note: We remove content that contains adult nudity or sexual activity.)
- Content that promotes the use of certain regulated products, such as tobacco or vaping products, adult products and services, or pharmaceutical drugs. (Note: We remove content that attempts to sell or trade most regulated goods.)
- Content that may promote or depict cosmetic procedures.
- Content that may be attempting to sell products or services based on health-related claims, such as promoting a supplement to help a person lose weight.

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.

Transparency & Accountability

We publish a quarterly Community Standards Enforcement Report that discloses metrics on the effectiveness of our policies and processes in reducing the prevalence of violent and graphic content on Facebook and Instagram.

The table below shows the pieces of content that we took action on globally in 2021 and the proactive rate of content detected before people reported it.[83]

Period	Facebook	Instagram
Jan-Mar	34 million with proactive rate over 99%	5.5 million with proactive rate over 98%
Apr-Jun	30.1 million with proactive rate over 99%	7.6 million with proactive rate over 98%
Jul-Sep	26.6 million with proactive rate over 99%	10.7 million with proactive rate over 99%
Oct-Dec	25.2 million with proactive rate over 99%	5.5 million with proactive rate over 98%

For New Zealand, in 2021:

- We took action on over 40 thousand pieces of content on Facebook in New Zealand for violating our Violent and Graphic Content policy. 99% of this content was detected proactively before people reported it to us.
- We took action on over 34 thousand pieces of content on Instagram in New Zealand for violating our Violent and Graphic Content policy. 98% of this content was detected proactively before people reported it to us.

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 general approach to online safety and harms).

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, 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

[83] https://transparency.fb.com/data/community-standards-enforcement/graphic-violence/facebook/

if something is untruthful — who or what is the source of truth — which often comes with differing views.

In 2021, we commissioned La Trobe University Academic Dr. Andrea Carson to study on misinformation regulation, where she notes: "The lack of universally agreed definitions of terms such as online misinformation, disinformation and fake news presents significant obstacles to achieving consensus on how to tackle the problem. Even among experts [...], significant diversity of opinion emerged over the meanings of misinformation and disinformation."[84]

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

Policies

Misinformation is different from other types of speech addressed in our Community Standards because there is no way to articulate a comprehensive list of what is prohibited. People have different levels of information about the world around them, and may believe something is true when it is not. A policy that simply prohibits "misinformation" would not provide useful notice to the people who use our services and would be unenforceable, as we don't have perfect access to information.

Our policies then articulate different categories of misinformation and try to provide clear guidance about how we treat that speech. Our approach reflects our attempt to balance our values of expression, safety, dignity, authenticity, and privacy. Under our Misinformation Policy,[85] we remove:

- Misinformation that is likely to contribute to a risk of imminent violence or physical harm
- Harmful health misinformation
- Voter or census interference
- Manipulated media

Oftentimes, misinformation can cut across different types of abuse areas: for example, a racial slur could be coupled with a false claim about a group of people which we would remove for violating our hate speech policy.

Our Repeat Violators policy allows us to take action on pages, groups, profiles, and accounts that repeatedly share misinformation or post content that violates our policies may — in addition to having their content removed — receive decreased distribution, be limited on their ability to advertise or monetise, be blocked from posting new content, or removed from our platforms altogether.

---

[84] A Carson, *Fighting Fake News: A Study of Online Misinformation Regulation in the Asia-Pacific*, https://www.latrobe.edu.au/__data/assets/pdf_file/0019/1203553/carson-fake-news.pdf

[85] https://transparency.fb.com/policies/community-standards/misinformation

As online and offline environments change and evolve, we will continue to consult with partners to evolve our misinformation policies.

Enforcement

Our approach to misinformation follows the three-part strategy — remove, reduce, and inform.[86]

- **Remove:** Our misinformation policy[87] allows us to remove misinformation content and pages, groups, profiles and accounts where it is likely to directly contribute to the risk of imminent physical harm. We also remove content that is likely to directly contribute to interference with the functioning of political processes and certain highly deceptive manipulated media (see section below on disinformation).
    - **Harmful misinformation.** In determining what constitutes misinformation in these categories, we partner with independent experts who possess knowledge and expertise to assess the truth of the content and whether it is likely to directly contribute to the risk of imminent harm. This includes, for instance, partnering with human rights organisations to determine the truth of a rumour about civil conflict. For example, we consult with the World Health organisation and public health authorities for guidance on our COVID-19[88] and harmful health misinformation policies.
    - **Ads including debunked or Community Standards-violating content.** Meta prohibits ads that include content debunked by third-party fact-checkers, as well as ads that violate our Community Standards. Advertisers that repeatedly post information deemed to be false or violate our Community Standards may have restrictions placed on their ability to advertise across Meta's platforms.
- **Reduce:** For misinformation that does not warrant removal, but nevertheless undermines the authenticity and integrity of our platform, we focus on reducing its prevalence. We know that people often use misinformation in harmless ways, such as to exaggerate a point or in humour or satire. They also may share their experience through stories that contain inaccuracies. In some cases, people share deeply-held personal opinions that others consider false, or share information that they believe to be true but others consider incomplete or misleading.
    - **Third-party fact-checking program.** In determining what is false, we rely on independent fact-checkers to review and rate the accuracy of stories - posts that have been rated false are demoted in Feed, which in turn reduces views to users. We also display strong warning labels and notify people who come across, try to share or already have shared the false-rated post. Based on one fact-check, we're able to kick off similarity detection methods that identify duplicates of debunked stories, applying the same enforcement penalties of reducing the distribution, showing warning labels, and notifying people.[89] Information on our fact-checking program can be found here.

[86] https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/
[87] https://transparency.fb.com/policies/community-standards/misinformation/
[88] https://www.facebook.com/help/230764881494641
[89] https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking

- Since 2016, our fact-checking program has expanded to include more than 90 organisations. The work of these fact-checkers have a global impact, as the treatment of their false-rated posts (i.e. demotion, notification and warning) are applied globally. As such, the posts will also be demoted for New Zealanders, who will also receive notifications and see warning labels if they have engaged with the posts. In New Zealand, specifically, Meta works with two independent fact-checking partners - the Agence France Presse (AFP) and the Australian Associated Press (AAP) who include New Zealand based, trained journalists in their team. Members of the public can also submit reports directly to fact-checkers and review debunking articles via the fact-checkers website.

- We have built the largest global fact-checking network of any platform and have contributed more than USD $100 million to programs supporting our fact-checking efforts since 2016. This includes direct support of fact-checkers for their work on our platforms as well as industry initiatives like sponsorships, fellowships, and grant programs. We also invest significant resources to support fact-checkers during moments of crises and war.[90]

- During the month of April, we put warning labels on about 50 million pieces of content related to COVID-19 on Facebook, based on around 7,500 articles by our independent fact-checking partners.[91]

- **Repeat violators.** We demote the content of pages, groups, profiles, and accounts that repeatedly share misinformation or post content that violates our Community Standards. Depending on the number and severity of the violations, these users may also lose their ability to advertise or monetise, be blocked from posting new content, or be removed from our platforms altogether.[92] It should be noted that this area is often highly adversarial as those who attempt to bypass our integrity measures, change their tactics in order to avoid enforcement. That is why we continue to update our policies and enforcement practices as new information and trends, both on and off platform, become available to us.

- **Content Distribution (Demotion) Guidelines.** Aside from fact-checked misinformation and content posted by repeat violators, we also demote other types of content, listed under our Content Distribution Guidelines[93], that are problematic and low quality content.

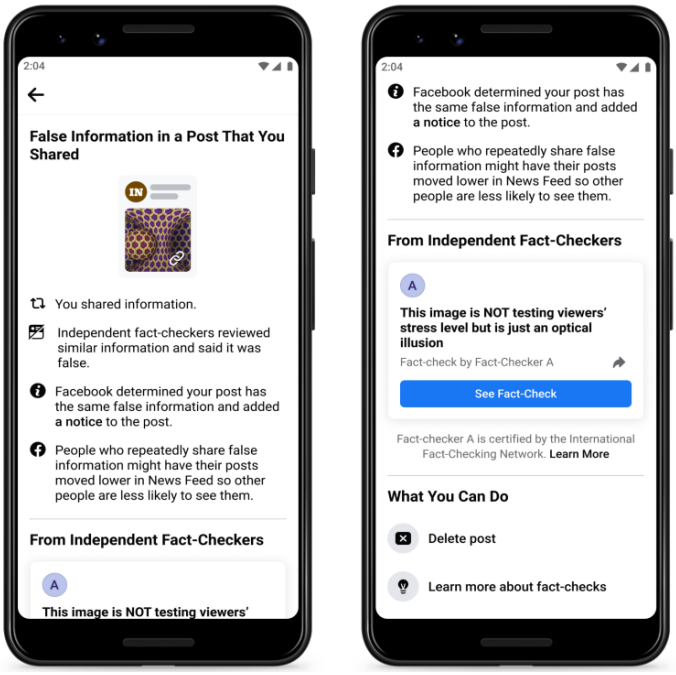
Tools, Products and Resources

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. Our strategy focuses on providing people with additional context and information on posts they

[90] https://www.facebook.com/formedia/blog/third-party-fact-checking-industry-investments
[91] https://about.fb.com/news/2020/04/covid-19-misinfo-update/.
[92] https://about.fb.com/news/2018/08/enforcing-our-community-standards/
[93] https://transparency.fb.com/features/approach-to-ranking/types-of-content-we-demote

see and connecting them with authoritative information. Some of the tools and products we have implemented include:

- **Warning labels.** Content across Facebook and Instagram that has been rated false or altered by our fact-checkers are prominently labelled so people can better decide for themselves what to read, trust, and share.

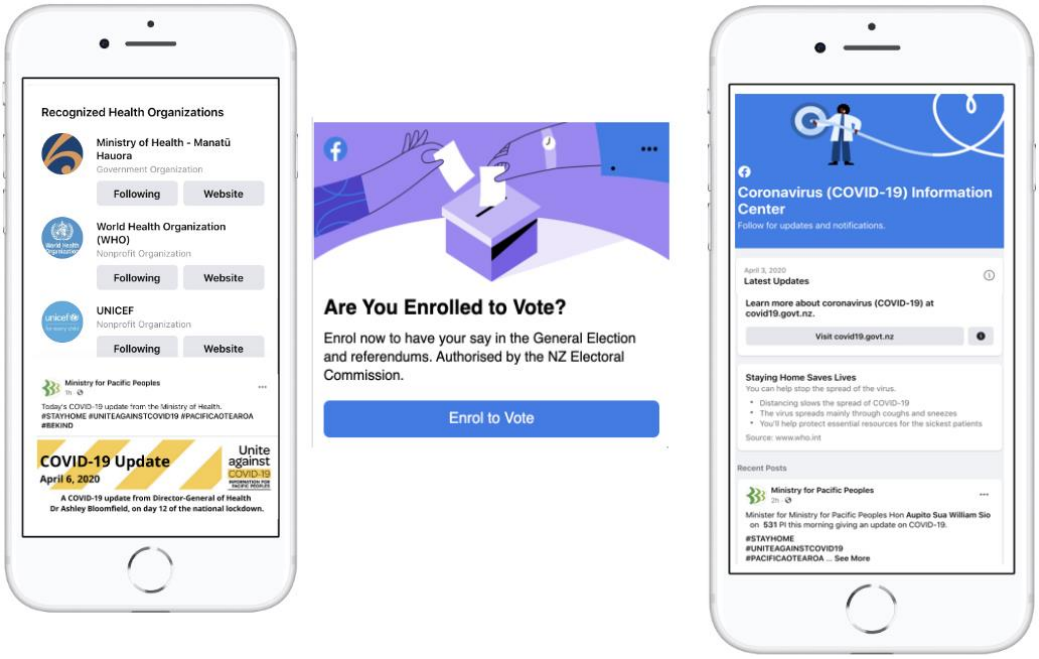


- **Connecting people to accurate and authoritative information.** We launch tools and products, such as Voter Registration and Election Day Reminders, to connect people with accurate information about when and how to vote. For COVID-19, we connect people with authoritative health sources through a number of methods, including directing people to the COVID-19 Information Center when they search for COVID-19 (see screenshot below). In New Zealand, the COVID-19 Information Center links to the Unite Against Covid-19, Ministry of Health, Ministry for Pacific Peoples resources in addition to global health resources. A similar effort has also been launched for climate change in which we direct people to our Climate Science

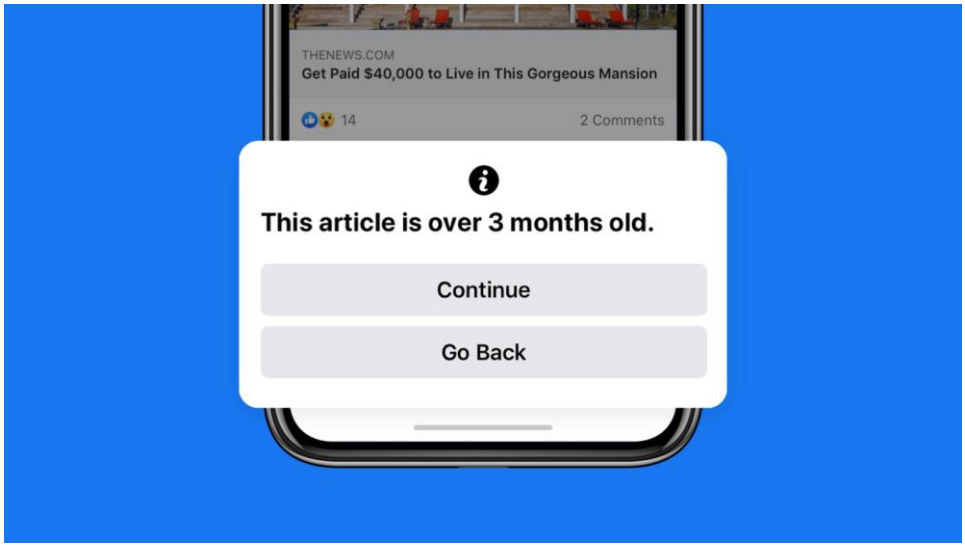
Information Center. These tools have connected hundreds of millions of people around the world to accurate and authoritative information.

- **More context about content people share.** Our goal is to make it easier for people to identify content that's timely, reliable and most valuable to them. As such, we introduced tools providing more context to help people decide what to read, trust and share:[94]

    - A notification screen lets people know when news articles they are about to share are more than 90 days old. People may continue sharing if they decide the article is relevant. News publishers, in particular, have expressed concerns

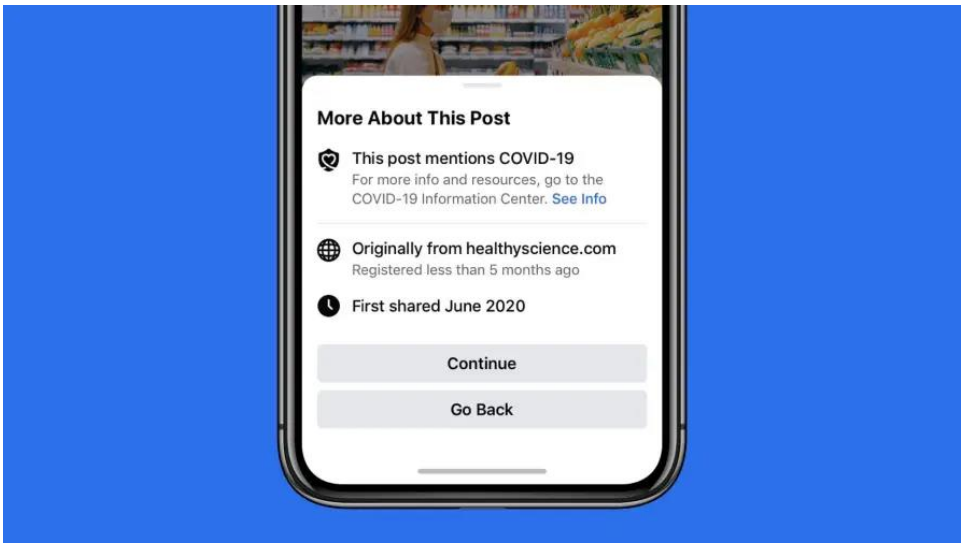


about older stories being shared on social media as current news, which can misconstrue the state of current events.

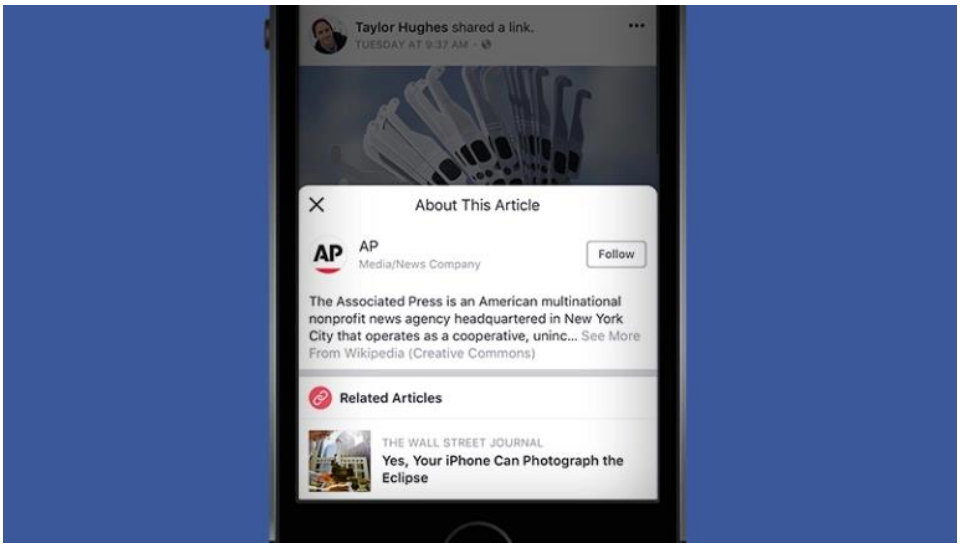


[94] https://about.fb.com/news/2020/06/more-context-for-news-articles-and-other-content/

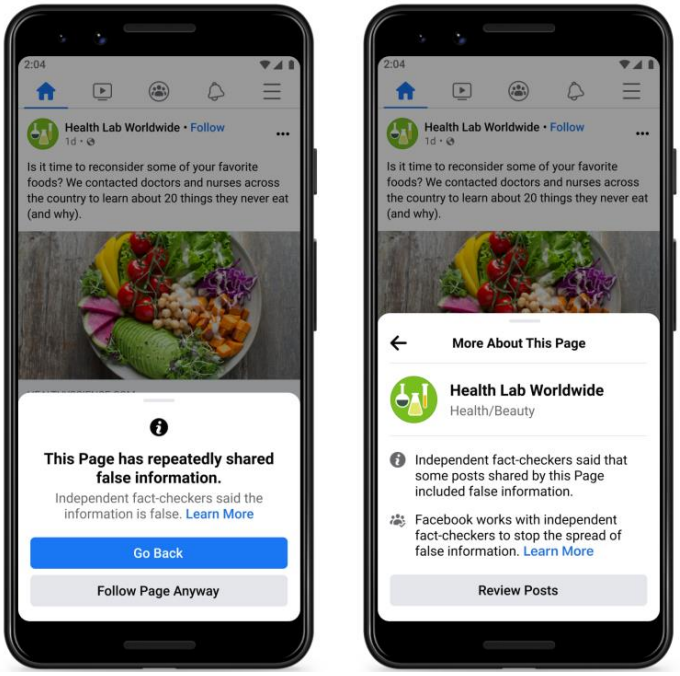
- A notification screen gives people more context about COVID-19 related links when they are about to share them. It also directs people to our COVID-19 Information Center to ensure people have access to credible information.



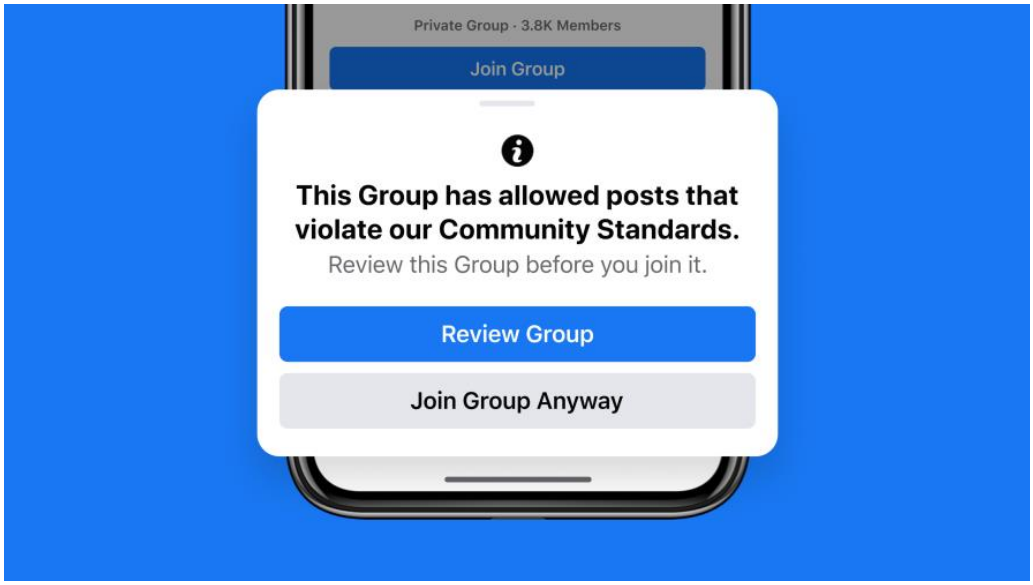
- **More context about news articles.** People can click on a button in the post to get contextual information pulled from across Facebook and other sources, such as information from the publisher's Wikipedia entry, a button to follow their Page, more articles from the publisher, and information about how the article is being shared by people on Facebook. In some cases, if that information is unavailable, we will let people know, which can also be helpful context.



- **Warning for Pages that repeatedly share false claims.** We warn people if they're about to "like" a Page that has shared misinformation to help users make informed decisions on whether they want to follow that Page. The warning includes links to additional contextual information, including a message that states the Page has shared false information. Repeated sharing of harmful false claims may ultimately lead to the removal of the Page on our platforms altogether.



- **Warning for Groups that repeatedly violate Community Standards.** We warn people if they're about to join a group that has Community Standards violations, so they can make a more informed decision before joining. We'll limit invite notifications for these groups, so people are less likely to join. For existing members, we'll reduce the distribution of that group's content so that it's shown lower in Feed.[95]



Partnerships

As noted, the third part of our approach to tackling misinformation is to **inform,** by providing people with accurate and authoritative information that will help them critically evaluate information, make informed decisions about the content they see, and self-correct when they have been exposed to misinformation.

[95] https://about.fb.com/news/2021/03/changes-to-keep-facebook-groups-safe/

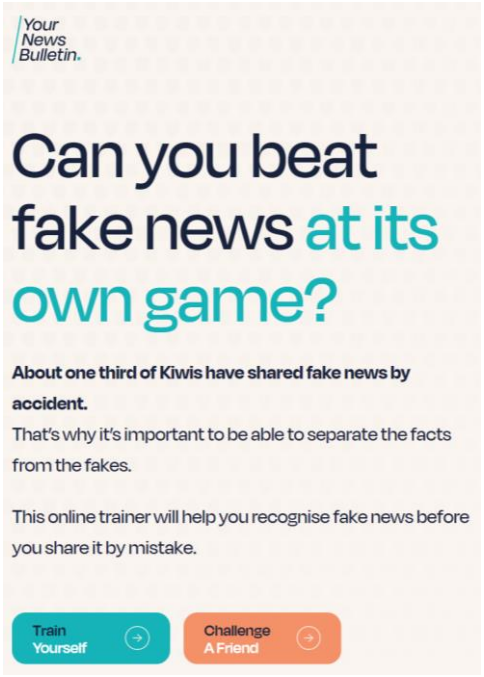
As noted by an international group of human rights experts (in relation to COVID-19): "it is essential that governments and internet companies address disinformation in the first instance by themselves providing reliable information… Resorting to other measures, such as content take-downs and censorship, may result in limiting access to important information for public health and should only be undertaken where they meet the standards of necessity and proportionality."[96]

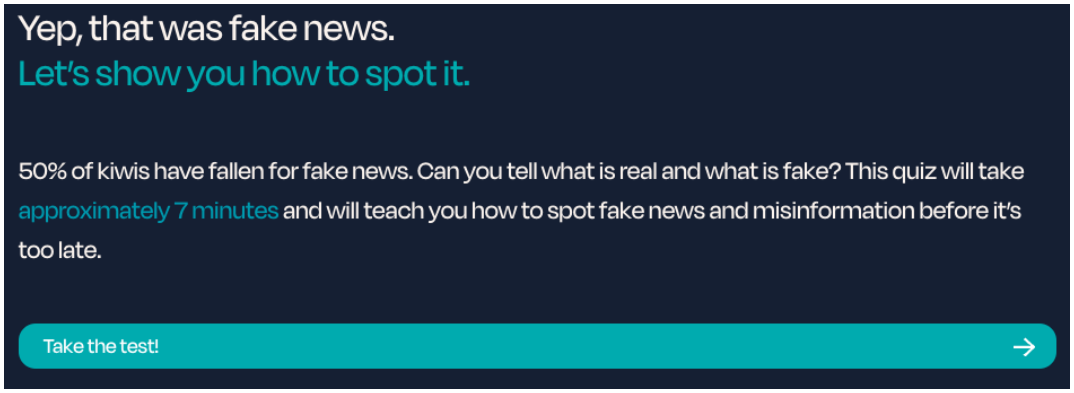
Combatting misinformation requires cross-sector collaboration. We continue to partner with industry, government, academics and civil society organisations to ensure the measures we take to address misinformation are based on expert information, and have the most effective impact, these initiatives include:

- In November 2021, under our partnership with misinformation experts First Draft (now the Information Futures Lab under Brown University), we launched a "Don't Be a Misinfluencer" campaign for public figures and creators. The campaign aimed to prevent the amplification of misinformation and includes a toolkit with information on how to identify and combat misinformation which was promoted by New Zealand influencers.[97] The project also included the 'Protect Your Voice' toolkit, which provides creators and other high profile account holders with resources to prevent the spread of misinformation on their own accounts, and to help amplify that message to their followers.

- Meta launched a media literacy campaign, 'Your News Bulletin', in partnership with Netsafe in 2020, which included an interactive activity and supporting resources to help develop media literacy skills. Over the course of the campaign in New Zealand, there were 180,000 unique users to the educational website and over 1.6 million people were reached through social media. The campaign has won several awards.

[96] https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25729
[97] First Draft, 'Protect your voice: a toolkit for Australian influencers and celebrities', *First Draft website,* https://firstdraftnews.org/tackling/protect-your-voice-a-toolkit-for-australian-influencers-and-celebrities/





- Meta has supported organisations with additional social media training, marketing and design support, and provided a significant amount of advertising credits. The type of support has varied across partners, but includes: Netsafe, Unite Against Covid-19, Ministry of Health, Te Puni Kōkiri (Karawhiua programme), Ministry for Pacific Peoples and UNICEF to promote authoritative health information on Covid-19 and the vaccine.
- Throughout 2021 we ran a number of campaigns through our services to combat misinformation, including through 'myth busting' messages to address the (evidence based) top myths about COVID-19 and vaccinations. We ran these campaigns in New Zealand, in English, Māori and Samoan. The campaign reached 1.8 million Kiwis and the material was viewed 10.7 million times.

- We have publicly provided a number of tools through our Data for Good programme that allows researchers and policymakers to follow trends in, for example, vaccine hesitancy, in order to better inform public messaging campaigns. We also onboarded a number of academics, including members of Te Pūnaha Matatini, to access data sets to assist in their modelling work, especially through our Social Connectivity Indexes.

Transparency & Accountability

Globally from January to June 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 2021:

- We removed over 24 thousand pieces of content on Facebook and Instagram in New Zealand violating our harmful health misinformation policy.
- We displayed warning labels on over 2.5 million distinct pieces of content on Facebook in New Zealand (including reshares) based on over 100 thousand articles written by our global third-party fact checking partners.

Disinformation

We actively work across our platforms to find and stop disinformation campaigns. But like the term "misinformation", "disinformation" is often used imprecisely and interchangeably to mean misinformation, foreign/electoral interference, influence operations, information manipulation and even defamation. At Meta, we try to bring clarity to the discussion on disinformation by using more specific terms, like "influence operations" (IO) and "coordinated inauthentic behaviour" (CIB), to describe coordinated efforts that aim to manipulate or corrupt public debate for a strategic goal.

For the purpose of this report, 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.

We take a three-prong approach to tackling disinformation — 1) preventing interference, 2) fighting misinformation, 3) increasing transparency.

We have grown the team focused on IO network disruptions to over 200 experts across the company, with backgrounds in law enforcement, national security, investigative journalism, cybersecurity, law, and engineering. We continue to build scaled solutions to help detect and prevent the proliferation of inauthentic accounts and behaviours, and we have partnered with civil society, researchers, and governments to strengthen our defences.

- **Preventing Interference.** Our goal is to prevent or stop IO actors from operating our platforms altogether.
    - We work with government authorities, law enforcement, security experts, civil society, and other tech companies to stop IO threats by establishing a direct line of communication, sharing knowledge and identifying opportunities for collaboration.
    - We continue to scale our investigations operations with people and tools in order to take down IO networks (i.e. inauthentic accounts, Pages and Groups) and identify emerging threats more quickly. We have removed tens of thousands of pages, groups and accounts involved in coordinated inauthentic behaviour – more than 50 networks just in 2021 alone.
    - We also continue to update our inauthentic behaviour policy, which covers **Coordinated Inauthentic Behaviour**, to improve our ability to counter new tactics and more quickly act against the spectrum of deceptive practices we see on our platforms - whether foreign or domestic, state or non-state.
- **Increasing Transparency.** We believe increased transparency leads to increased scrutiny and helps people better understand why they see the content they see and who is behind them. We have introduced tools and products to bring greater transparency around political advertising, Pages and posts, so that people can see who is trying to influence them. We also regularly publish reports on our CIB efforts.[98]

[98] about.fb.com/news/tag/coordinated-inauthentic-behavior/

Policies

Two key markers for influence operations (IO) are inauthenticity and coordination. Actors engaged in IO need not necessarily use misinformation; most of the content shared by IO campaigns are not provably false, and would in fact be acceptable political discourse if it was shared by authentic actors. The real issue is that the actors behind these campaigns are using deceptive behaviours to conceal the identity of the organisation behind a campaign, make the organisation or its activity appear more popular or trustworthy than it is, or evade enforcement efforts.

Our Inauthentic Behaviour[99] policy is targeted at addressing these deceptive behaviours. In line with our commitment to authenticity, we do not allow people to misrepresent themselves on Facebook, use fake accounts, artificially boost the popularity of content or engage in behaviours designed to enable other violations under our Community Standards.

Under this Policy, is our policy on Coordinated Inauthentic Behavior (CIB) that aims to address IO campaigns directly. Defined as "the use of multiple Facebook or Instagram assets, working in concert to engage in Inauthentic Behavior (as defined by our policy), where the use of fake accounts is central to the operation", the policy informs how we find, identify and remove IO networks on our platforms.

Working alongside the CIB policy is our policy on Account Integrity and Authentic Identity, which allows us to remove millions of fake accounts every day. Our goal is to remove as many fake accounts on Facebook as we can to minimise opportunities for IO 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.

Enforcement

Our approach to Coordinated Inauthentic Behavior (CIB) more broadly, is grounded on behaviour- and actor-based enforcement. This means that we are looking for specific violating behaviours exhibited by violating actors, rather than violating content (which is predicated on other specific violations of our Community Standards, such as misinformation and hate speech). Therefore, when CIB networks are taken down, it is based on their behaviour, not the content they posted.

We try to stop fake accounts abusing our platforms in three distinct ways:[100]

- **Blocking accounts from being created.** Our systems look for a number of different signals that indicate if accounts are created en masse from one location. A simple example is blocking certain IP addresses altogether so that they can't access our systems and thus can't create accounts.
- **Removing accounts when they sign-up.** We try to spot signs of malicious behaviour through a combination of signals such as patterns of using suspicious email addresses, suspicious actions, or other signals previously associated with other fake

[99] https://transparency.fb.com/policies/community-standards/inauthentic-behavior/
[100] https://about.fb.com/news/2019/05/fake-accounts/

accounts we've removed. Most of the accounts we currently remove are blocked within minutes of their creation before they can do any harm.

- **Removing existing accounts.** Some accounts may get past the above two defences and still make it onto the platform. Often, this is because they don't readily show signals of being fake or malicious at first. We find these accounts when our detection systems identify inauthentic behaviour or if users report them to us. We use a number of signals about how the account was created and is being used to determine whether it has a high probability of being fake and disable those that are.

Pages and Groups directly involved in CIB activity may also be removed when detected as part of the network. In turn, posts published by these accounts would be taken down. Taking this actor- and behaviour-based approach essentially allows us to address the problem at the source.

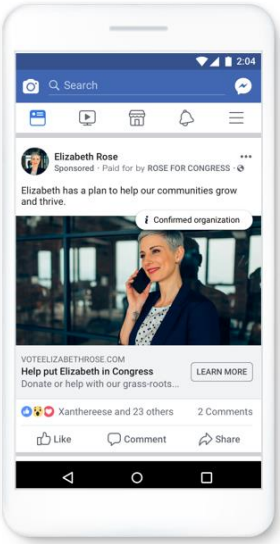
After each takedown, we feed the data about the network into our automated detection systems to block the network from operating on our platforms again, as well as explore ways to make our platforms more resilient and difficult to exploit. Using both automated and manual detection, we continuously remove accounts, Pages and Groups connected to networks we took down in the past.

For a comprehensive overview of our approach, see here.

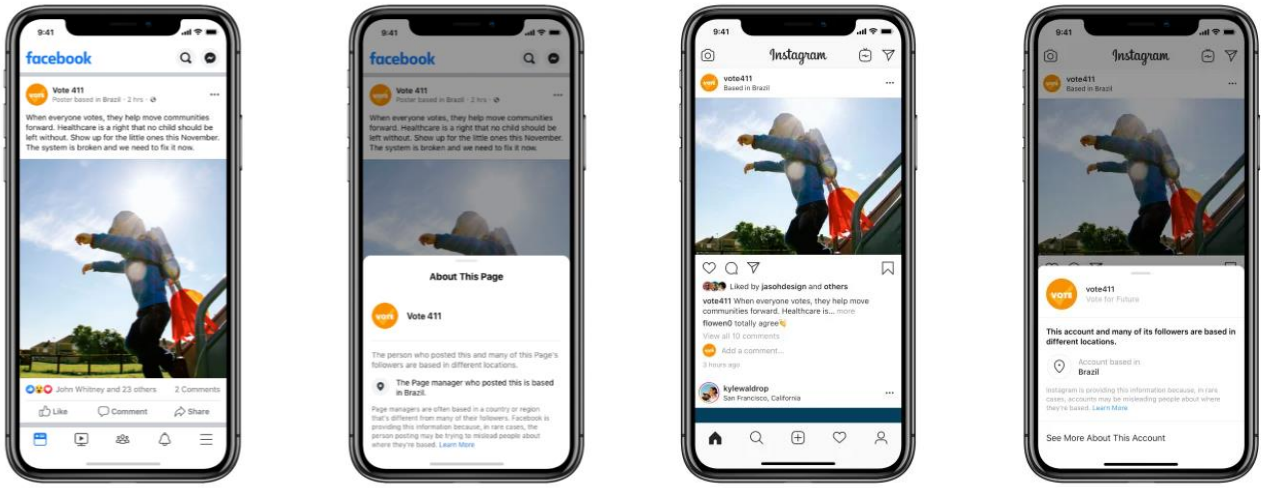
Tools, Products and Resources

We have tools to bring greater transparency around political advertising, Pages and posts, so that people can see who is trying to influence them. We have also introduced tools that give users more control over the content they see. Specifically, we have introduced:

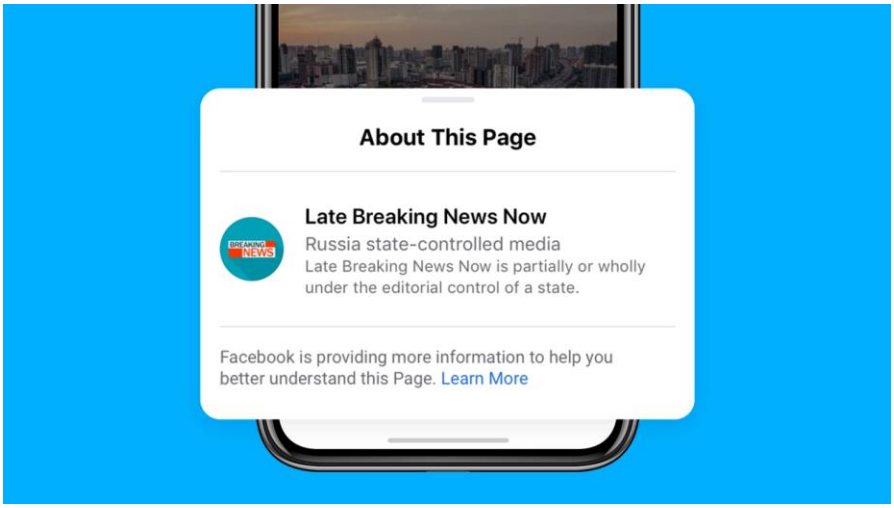
- **Political ads transparency tools.** Advertisers running ads about social issues, elections or politics are required to go through our authorization process (which includes proving who they are and where they live) and apply a "Paid for by" disclaimer label to show who's behind the ad. These ads are then housed in a public searchable Ad Library for seven years. Information in the Ad Library includes spend, demographic, and targeting data about an ad, as well as information about the advertiser (this feature has been running in New Zealand since before the 2020 general election).



- **Page Transparency.** We show information about a Page, such as when it was created, name changes, and the location(s) of the Page admins. For Instagram accounts with large audiences, people can see information such as the country where the account is located.



- **State-Controlled Media Label.** We want to help people better understand who's behind the news they see, so we label media outlets that we believe are wholly or partially under the editorial control of their government as state-controlled media.



Partnerships

We work with government authorities, law enforcement, security experts, civil society and other tech companies around the world to stop IO threats by establishing a direct line of communication, sharing knowledge and identifying opportunities for collaboration. Information-sharing enables Meta, investigative journalists, government officials, academia and industry peers to better understand and expose internet-wide security risks.

We share information on the networks with third-party independent researchers, such as the Atlantic Council's DFRLab, Stanford's Internet Observatory and Graphika (who publish their own reports). When appropriate, we also share what we know with relevant law enforcement, security experts, civil society, researchers and industry partners so they can take appropriate action. This includes appropriate New Zealand authorities .

Transparency & Accountability

We publish CIB reports to raise awareness of IO threats on our platforms and show the progress we are making. These reports also allow researchers, journalists, policymakers, and security experts scrutinise our work.

- Since 2017, our security teams at Facebook have reported on over 150 covert influence operations for violating our policy against Coordinated Inauthentic Behavior (CIB). These operations targeted public debate across both established and emerging social media platforms, as well as local blogs and major newspapers and magazines. They were foreign and domestic, run by governments, commercial entities, politicians, and conspiracy and fringe political groups.[101]

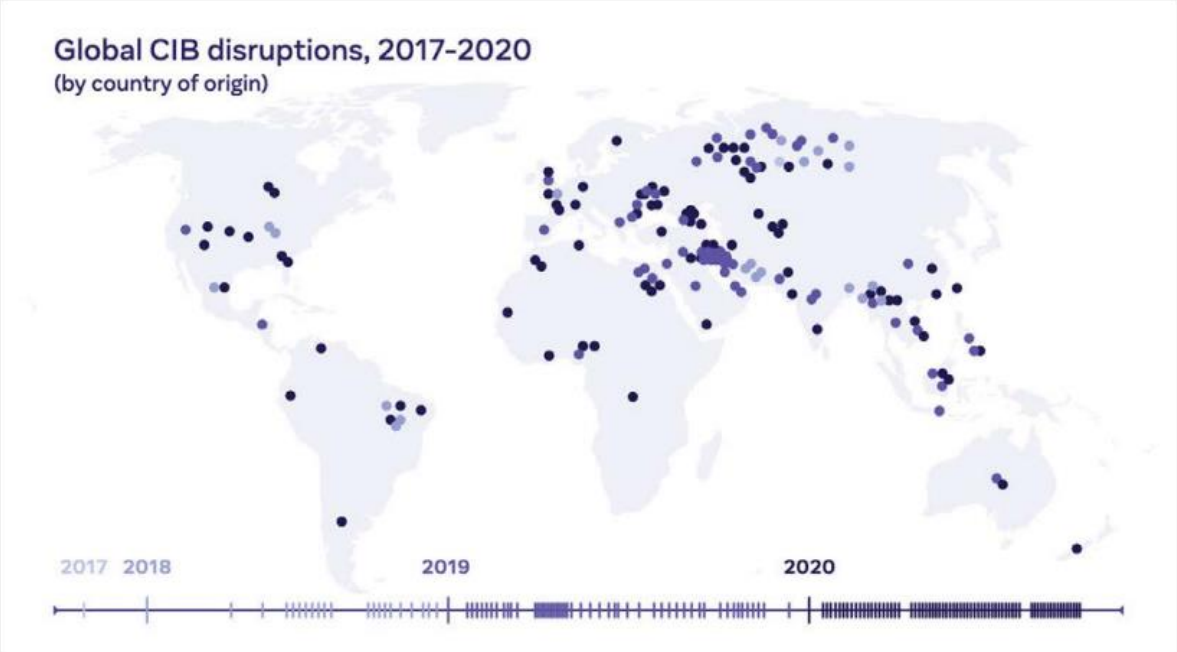
- In 2021, we removed 52 networks that engaged in coordinated efforts to manipulate or corrupt public debate for a strategic goal, while relying centrally on fake accounts to mislead people about who's behind them. They came from 34 countries, including Latin America, the Asia-Pacific region, Europe, Middle East and Africa.

- We have taken action on instances of CIB operations that New Zealanders. In 2020, we removed an operation that operated from many regions around the world including the US, Canada, Australia, New Zealand, Vietnam, Taiwan, Hong Kong, Indonesia, Germany, the UK, Finland and France. It targeted primarily English and Chinese-speaking audiences globally and Vietnam. Our investigation linked this network to Truthmedia, a digital media outlet, which is now banned from our platforms.[102]

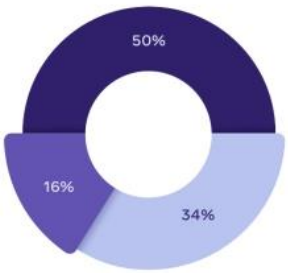
- In 2021, we published our first State of Influence Operations report that recapped our CIB efforts from 2017-2021. The threat report drew on our existing public disclosures and our internal threat analysis to do four things: 1) defines how CIB manifests on our platform and beyond; 2) analyses the latest adversarial trends; 3) uses the US 2020 elections to examine how threat actors adapted in response to better detection and enforcement; and 4) offers mitigation strategies that we've seen to be effective against IO.

[101] https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf https://about.fb.com/wp-content/uploads/2022/01/December-2021-Coordinated-Inauthentic-Behavior-Report-2.pdf

[102] https://about.fb.com/news/2020/08/july-2020-cib-report/



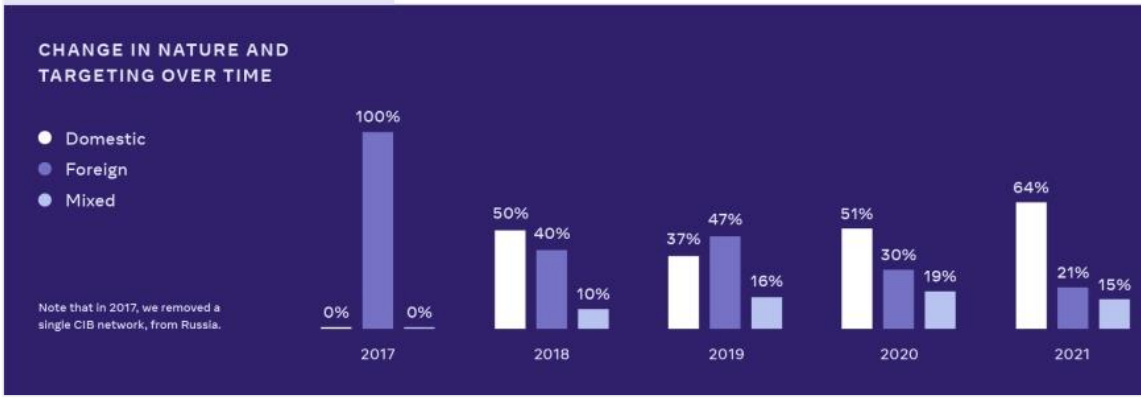
Nature of Coordinated Inauthentic Behavior networks we disrupted



TARGET AUDIENCES (2017-2021)

- Domestic (home country)
- Foreign (countries abroad)
- Mixed (both home and abroad)

CHANGE IN NATURE AND TARGETING OVER TIME



Countries most frequently targeted by influence operations, 2017-2020

In addition to our CIB specific transparency disclosures, we also disclose metrics on fake account removals in our Community Standards Enforcement Report. As noted above, fake accounts is a key part of our strategy to keep IO threat actors off our platforms. The table below shows the number of accounts removed globally in 2021 and the proactive rate of fake accounts detected before people reported them.[103]

Period	Facebook	Instagram
Jan-Mar	1.3 billion with proactive rate over 99%	not available
Apr-Jun	1.7 billion with proactive rate over 99%	not available
Jul-Sep	1.8 billion with proactive rate over 99%	not available
Oct-Dec	1.7 billion with proactive rate over 99%	not available

4. Empower users to have more control and make informed choices

As required by section 4.2 of the Code, Meta is committed to empowering users to have greater control and be better informed over the content they see and/or their experiences and interactions online.

We believe that 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

[103] https://transparency.fb.com/data/community-standards-enforcement/fake-accounts/facebook/

this by providing greater transparency and control to users; providing information that will help them make informed decisions; and advancing media and digital literacy.

Throughout this report, we have outlined many of the tools, products and resources we have introduced to users to address the different areas of harms. Each tool or resource serves a different purpose or solves a different problem. This includes:

- 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 allow users to decide how they want content ranked

5. Enhance transparency of policies, processes and systems

As required by section 4.3 of the Code, Meta is committed to making transparent our safety and integrity-related policies, processes and systems, i.e. 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.

We have 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[104], Help Centers (Facebook, Instagram) or Newsroom.

To build on what has been shared in the sections above, this section provides a breakdown of transparency reports we regularly publish to provide further insight into our online safety and content moderation practices.

At Meta, 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. While our initial reports focused on the nature and extent of government requests for user data, 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.

- **Community Standards Report.**[105] We publish the Community Standards Enforcement Report on a quarterly basis to track our progress and demonstrate our continued commitment to making Facebook and Instagram safe and inclusive.

[104] https://transparency.fb.com/
[105] https://transparency.fb.com/data/community-standards-enforcement/

- **Content restrictions.**[106] 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.**[107] 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.**[108] 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.**[109] 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.

- **Adversarial Threat Report.** We publish a quarterly adversarial threat report[110] 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 Behaviour (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.

- **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

[106] https://transparency.fb.com/data/content-restrictions/
[107] https://transparency.fb.com/data/government-data-requests/
[108] https://transparency.fb.com/data/internet-disruptions/
[109] https://transparency.fb.com/data/intellectual-property/
[110] https://about.fb.com/?s=adversarial+threat+report

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.

6. Support independent research and evaluation

As required by section 4.4 of 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

We support independent research to solve some of the world's greatest challenges. We are also deeply committed to protecting our users' privacy and maintaining a safe and secure community. We have worked to promote research - while preserving privacy - through multiple initiatives. Our investments in independent research is an inherent part of our overall efforts to make the internet and people on our platforms more secure. It helps us develop a foundational understanding of how best to serve our community — by building better products and offering valuable services — and deepens our understanding of the impact our products and services may have on society.

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. We currently offer 3 efforts to support research:

    1. **Ad Targeting Transparency Data Sets**, which includes detailed targeting information for social issue, electoral and political ads that ran globally since August 2020. This data is provided for each individual ad and includes information like the interest categories chosen by advertisers. We built the Meta Researcher Platform to enable qualified academic researchers to study social media's impact on society.

    2. **URL Shares Data Set**, which includes differentially private individual-level counts of the number of people who viewed, clicked, liked, commented, shared, or reacted to any URL (for any URL with at least 100 public shares) on Facebook between January 2017 and July 2019. Counts are aggregated at the level of country, year-month, age bracket, gender, and for U.S. users, political page affinity. The URL Shares data set is regularly updated to add additional year-months and countries. In order to maintain the independence

of researchers who use these data, access to the URL Shares is granted by Social Science One. New researchers are onboarded once per quarter and access is governed by a Research Data Agreement.

3. **Researcher API.** In 2021, we piloted the Researcher API that gives qualifying academics access to near real-time data, as well as billions of historical data points. The API was specifically designed for academic needs and allows them to conduct longitudinal research across all public Facebook Pages, Groups, Posts and Events in the US and select EU countries. Researchers can use the API to understand how public discussions on Facebook influence the social issues of the day. We offer this product via the Researcher Platform, which allows us to share privacy-protected data in a secure way. We have invited a small group of qualified academics to test this product and provide feedback so we can iterate and improve it, before launching to a broader group of researchers.

- **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. The program builds maps, surveys, and insights (with the use of privacy-preserving data) to help strengthen communities and advance social issues. Data for Good tools can, for example, help organisations better respond to disease outbreaks. During the COVID0-19 pandemic, for example, Data for Good tools were used by health authorities around the world to plan vaccination campaigns. In partnership with Yale University, earlier this year, we launched a survey to gather public views towards climate change around the globe, with the aim of informing policy decisions and priorities for governments, as well as a resource to inform research and awareness raising campaigns by activists and NGOs.

- **Research Platform for CIB Network Disruptions.** Since 2018, we have been sharing information with independent researchers about our network disruptions relating to Coordinated Inauthentic Behaviour (CIB), and in 2021, we expanded our beta research platform — with about 100 data sets — to more researchers studying influence operations worldwide. This platform provides access to raw data where researchers can visualise and assess these network operations both quantitatively and qualitatively. In addition, we share our own internal research and analysis.

- **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. We also look for research that will help develop a foundational understanding of how best to serve our community and contribute to our understanding of societal trends. The following are some of the key research grants and awards we have supported for this purpose. Information on the Meta Research team and program can be found here.

    - **Foundational Integrity Research on Misinformation and Polarisation.** Launched in 2020, the aim of this research program is to support the growth of scientific knowledge in the areas of misinformation, polarisation, information quality, and social conflict on social media and social technology platforms. It also aims to contribute to a shared understanding across the broader scientific community and technology industry on how social technology companies can better address social issues on their platforms. In

2020, we funded research led by the University of Canterbury concerning the impacts on polarisation of online interventions. In 2021, we provided $1,000,000 USD of funding to researchers in 21 countries, including New Zealand. Winning research proposals include studies on the over-time effects of indirect exposure to misinformation; testing fact and logic-based responses to polarising climate misinformation; and media literacy intervention to debunk out-of-context visual posts. The full list of winners can be found here.[111]

- **Instagram Research Awards on Safety and Community Health.** Last May, we launched an external research initiative focused on safety and community health, especially as it relates to young people and underserved communities, with the aim of helping us: (1) better understand equity and fairness issues in our community, (2) develop better policies, (3) assess possible improvements to protect our younger community, or (4) better understand the mechanisms (e.g., social support, social comparison) through which Instagram usage could impact the people that use our service. Award winners include proposals on mitigating cyberbullying experiences of younger users; chatbots as social support actors; and proactive moderation of coordinated harassment.[112]

- **Research on CSAM Sharers.** In 2021, to inform the development of tools targeted at reducing the sharing of child exploitative content, we consulted the world's leading experts (including the National Center for Missing and Exploited Children) and their research on child exploitation. The aim was to develop a research-backed taxonomy to categorise a person's apparent intent in sharing such content, which then allowed us to develop and test new tools. From this research, we launched a pop-up that is shown when people search for terms on our platforms 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. We also launched a safety alert that informs people who have shared viral, meme child exploitative content about the harm it can cause, our related policies and legal consequences for sharing this material.[113]

- **Study of Online Misinformation Regulation in the Asia Pacific.** In 2021, we commissioned independent research by respected Australian academic Dr Andrea Carson to map government approaches to combatting misinformation around the world, focussing on the Asia-Pacific region.[114] The report 'Fighting Fake News' has helped inform policymakers' thinking on approaches to misinformation and regulation.

- **Disinformation and Misinformation Amongst Diaspora Groups.** In 2021, we provided support for an analytical paper by First Draft on disinformation and

[111] https://research.facebook.com/blog/2021/9/announcing-the-2021-recipients-of-research-awards-in-misinformation-and-polarization/
[112] https://research.facebook.com/blog/2021/12/announcing-the-recipients-of-instagram-research-awards-on-safety-and-community-health/
[113] https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/
[114] A Carson, '*Fighting Fake News: A Study of Online Misinformation Regulation in the Asia-Pacific*', https://www.latrobe.edu.au/__data/assets/pdf_file/0019/1203553/carson-fake-news.pdf

misinformation amongst diaspora groups with a focus on Chinese language.[115] The paper aimed to inform policymakers on how to reduce misinformation within Chinese diaspora communities ahead of the federal election in Australia, which has attracted interest by policymakers and academia in other countries.

- **Fellowship Program.** Launched in 2010, our Fellowship Program[116] aims to foster ties with the academic community and support the research of promising doctoral students who are engaged in innovative and relevant research in areas related to computer science and engineering at an accredited university. We have funded over 200 students from around the world since the Fellowship program's conception.

Independent Evaluation

Meta is a founding member and signatory of the Aotearoa New Zealand Code of Practice for Online Safety and Harms and is committed to supporting independent reviews of the annual compliance reports submitted by Meta.

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.

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.

We believe independent evaluation is important to hold companies like Meta accountable and help us do better.

Note:

1. The New Zealand specific metrics in this report are our best estimates of content we took action on and of proactive rates based on the creator of the content and predicted country locations for those users.

2. The New Zealand specific metrics in this report are based on definitions and caveats as disclosed in Content Actioned and Proactive Rate.

3. Given that such violations are also highly adversarial, country-level data may be less reliable. For example, bad actors may often try to avoid detection by our systems by masking the country they are coming from. While our enforcement systems are global and will try to account for such behaviour, this makes it very difficult to attribute and report the accounts or content by producer country (where the person

[115] E Chan, S Zhang, 'Disinformation, stigma and chinese diaspora: policy guidance for Australia', *First Draft website,* 31 August 2021, https://firstdraftnews.org/long-form-article/disinformation-stigma-and-chinese-diaspora-policy-guidance-for-australia/

[116] https://research.fb.com/programs/fellowship/

who posted content was located). Given the global nature of our platforms where content posted in one country may be viewed almost anywhere across the world, other ways to attribute the country of content removed in a technically feasible and repeatable manner, become almost meaningless. So these estimates should be understood as directional best estimates of the metrics.

4. This report shares metrics in which we have considerable confidence in their accuracy across Facebook and Instagram. As we develop metrics for new policy areas, we will continue to expand this report, similar to our Community Standards Enforcement Report.