Meta's response to the Aotearoa New Zealand Code of Practice for Online Safety and Harms | October 2024

Signatory:	**Meta Platforms, Inc.** • Meta's mission is to give people the power to build community and bring the world closer together. We build technology that helps people connect, find communities, and grow businesses. We help people discover and learn about what is going on in the world around them, enable people to share their opinions, ideas, photos and videos, and other activities with audiences ranging from their closest family members and friends to the public at large, and stay connected everywhere by accessing our products.

Relevant Products / Services:	**Facebook & Instagram**

Executive Summary:	Meta is proud to be a founding member and signatory of the Aotearoa New Zealand Code of Practice for Online Safety and Harms ("the Code"). This industry Code is an important initiative that aims to encourage collaboration between the technology industry, civil society and government to combat online harms in a way that respects freedom of expression. This is our third annual Transparency Report submitted under the Code and outlines Meta's new developments in ensuring online safety and combatting online harm in New Zealand over the 12-month period from July 2023 to June 2024. Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta globally and in New Zealand, and have been published on the Code website. Some highlights of our work outlined in this report include: • Introducing enhanced product features and user controls, as well as product features to empower youth, parents and guardians to control their experiences and interactions online;

	• Launching new tools, technology and partnerships to support independent research and understand the impact of safety interventions and harmful content on society; • Supporting local non-governmental organisation partners in New Zealand to deliver youth and online safety initiatives; • Disrupting co-ordinated inauthentic behaviour by foreign interference networks that impacted New Zealand, among other countries; and • Combatting misinformation ahead of New Zealand's October 2023 election. We look forward to continuing to work with New Zealand policymakers, civil society, academics and experts on steps to combat harmful content online over the next year.

4.1 Reduce the prevalence of harmful content online

Signatories commit to implementing policies, processes, products and/or programs that would promote safety and mitigate risks that may arise from the propagation of harmful content online, as it relates to the themes identified in section 1.4.

Outcome 1. Provide safeguards to reduce the risk of harm arising from online child sexual exploitation & abuse (CSEA)
Measure 1. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to prevent known child sexual abuse material from being made available to users or accessible on their platforms and services Measure 2. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to prevent search results from surfacing child sexual abuse material Measure 3. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to adopt enhanced safety measures to protect children online from peers or adults seeking to engage in harmful sexual activity with children (e.g. online grooming and predatory behaviour) Measure 4. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to reduce new and ongoing opportunities for the sexual abuse or exploitation of children Measure 5. Work to collaborate across industry and with other relevant stakeholders to respond to evolving threats
Meta is committed to ensuring the safety and well-being of youth and children on its platforms. We have implemented a zero-tolerance policy towards child sexual exploitation and abuse, and are constantly working to improve and develop new technologies to prevent and detect such harms.

We've spent a decade working on this issue and continue to use a rights-based approach for the design of our services and content policy. Our Best Interests of the Child framework is informed by the UN Convention on the Rights of the Child, as well as regulation and guidance such as the UK Age Appropriate Design Code, the Irish Data Protection Commission's Children's Fundamentals and the French Commission Nationale de l'Informatique et des Libertés CNIL) Recommendation on Minors.

Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to combat child exploitation, globally and in New Zealand.

Our latest efforts to combat child exploitation include:

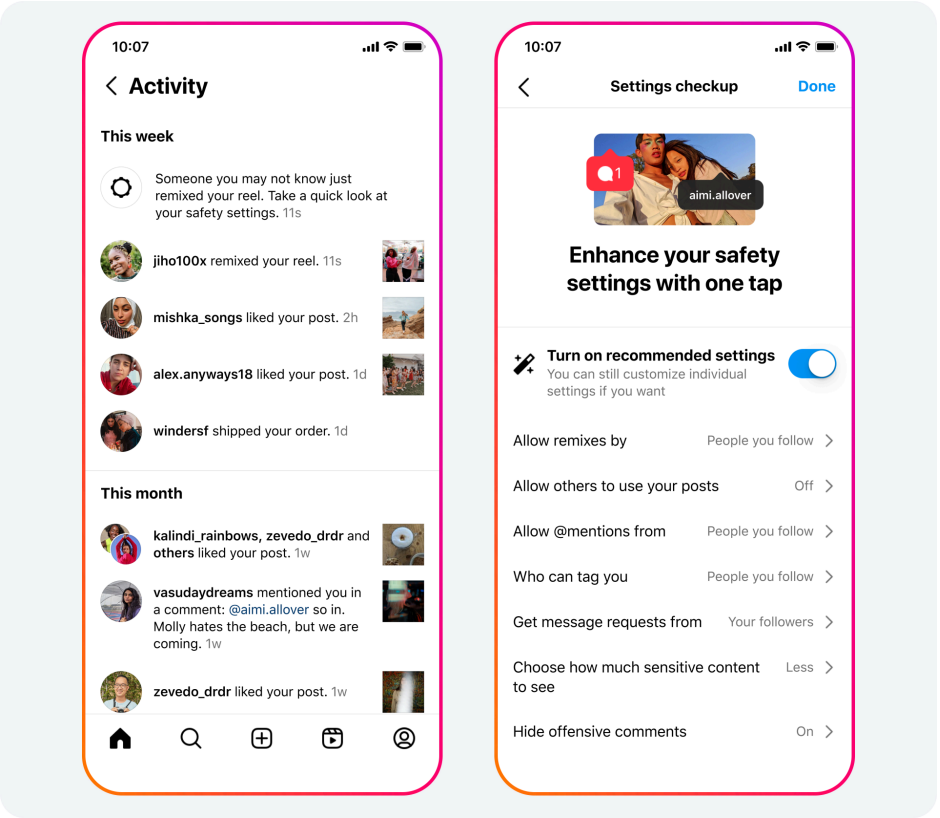
- Implementing new tools and continuously updating our policies to prevent the exploitation and abuse of children,
- Detecting, removing, and reporting any exploitative activity that violates our policies, and
- Collaborating with experts and authorities to ensure the safety of children on our platforms

Giving teens more age-appropriate experiences

- Meta is dedicated to providing teens with secure and age-appropriate experiences on its platforms. To support teens and their parents over the past decade, Meta has developed more than 30 tools and resources. These policies and efforts are aimed at addressing content that violates Meta's rules or could be considered sensitive.

Prompting teens to easily update their privacy settings

- To promote regular safety and privacy checks among teens on Instagram, we're introducing new notifications that encourage them to easily update their settings to a more private experience with just one tap. If they opt to "Turn on recommended settings", we will automatically adjust their settings to limit who can repost their content, tag or mention them, or include their content in Reels Remixes. Additionally, we'll ensure that only their followers can message them and help hide offensive comments, providing a safer and more controlled environment for teens.



Updates to our work on fighting online predators

- Meta's Child Safety Task Force has been working on three key areas to enhance the safety of children on our platforms: Recommendations and Discovery, Restricting Potential Predators and Removing Their Networks, and Strengthening Our Enforcement.
- In December 2023, we updated our policies to automatically disable accounts that exhibit a certain number of suspicious behaviour signals, which we monitor through over 60 different indicators. These signals include actions such as a teen blocking or reporting an adult, or someone repeatedly searching for terms that may suggest suspicious behaviour. We use this technology to limit potentially suspicious adults from interacting with teens and have expanded it to prevent them from interacting with each other as well.
- Additionally, we have expanded the list of child safety-related terms, phrases, and emojis that our systems can detect. This list is compiled from various sources, including non-profit organisations, online safety experts, our specialist child safety teams who investigate predatory networks, and our own technology that identifies misspellings or spelling variations of these terms. By continuously updating and refining our detection methods, we aim to provide a safer environment for children on our platforms.

Participation in the Lantern program

- In November 2023, Meta became a founding member of the Lantern program, a collaborative effort among technology companies to share signals about accounts and behaviours that violate child safety policies. Meta provided the technical infrastructure for the program and encouraged industry partners to participate. We manage and oversee the technology in partnership with the Tech Coalition, ensuring its ease of use and effectiveness

in providing partners with the necessary information to track down potential predators on their own platforms.

- Lantern enables participating companies to share various signals about accounts and behaviours that violate child safety policies, allowing them to conduct investigations and take appropriate action on their own platforms.
- A notable example of Lantern's value is a case study from the program's pilot phase, where MEGA, a Lantern partner, shared URLs with the program that they had previously removed for violating their child safety policies. Meta's specialist child safety team used this information to conduct a wider investigation into potentially violating behaviours related to these URLs on our platforms. As a result, over 10,000 violating Facebook Profiles, Pages, and Instagram accounts were removed. In accordance with legal obligations, we reported the violating profiles, pages, and accounts to the National Center for Missing and Exploited Children (NCMEC). Additionally, we shared details of our investigation back to Lantern, enabling participating companies to use the signals to conduct their own investigations.

Meta partners with New Zealand organisations and experts to support community initiatives and research relating to combatting child sexual exploitation and abuse. Some of our most recent efforts include:

- **Safeguarding Children New Zealand:** In September 2023, Meta partnered with Safeguarding Children NZ to host the "Child Safeguarding Week", focusing on preventing child abuse and neglect in online environments. Meta provided financial and advertising support for this week-long initiative. As part of the campaign, e-learning courses were offered to educate children on recognising, responding to, and ideally preventing opportunities for sexual abuse in their lives. The "Recognising and Responding to Grooming" course aimed to teach participants about the grooming process, enabling them to prevent and protect children from experiencing abuse and understand the harmful effects of grooming on children.

Global metrics for child endangerment content that we took action on globally in 2023 and the proactive rate of content we detected before people reported it.

Period	Child Nudity and Physical Abuse	Child Sexual Exploitation
Jan-Mar	Facebook: 1.9 million with proactive rate over 98% Instagram: 567,000 with proactive rate over 97%	Facebook: 8.9 million with proactive rate over 98% Instagram: 8.7 million with proactive rate over 99%
Apr-Jun	Facebook: 1.7 million with proactive rate over 97% Instagram: 321,000 with proactive rate over 96%	Facebook: 7.2 million with proactive rate over 96% Instagram: 1.7 million with proactive rate over 97%

Jul-Sep	Facebook: 1.8 million with proactive rate over 98% Instagram: 228,000 with proactive rate over 96%	Facebook: 16.9 million with proactive rate over 99% Instagram: 1.6 million with proactive rate over 96%
Oct-Dec	Facebook: 1.9 million with proactive rate over 99% Instagram: 199,000 with proactive rate over 95%	Facebook: 16.2 million with proactive rate over 96% Instagram: 2.1 million with proactive rate over 96%

For New Zealand, in 2023:

- **We took action on over 5,000 pieces of content on Facebook in New Zealand for violating our Child Nudity and Physical Abuse policies.** Over 97% of this content was detected proactively before people reported it to us.
- **We took action on over 4,000 pieces of content on Instagram in New Zealand for violating our Child Nudity and Physical Abuse policies.** Over 99% of this content was detected proactively before people reported it to us.
- **We took action on over 17,000 pieces of content on Facebook in New Zealand for violating our Child Sexual Exploitation policies.** Over 99% of this content was detected proactively before people reported it to us.
- **We took action on over 4,000 pieces of content on Instagram in New Zealand for violating our Child Sexual Exploitation policies.** Over 96% of this content was detected proactively before people reported it to us.

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

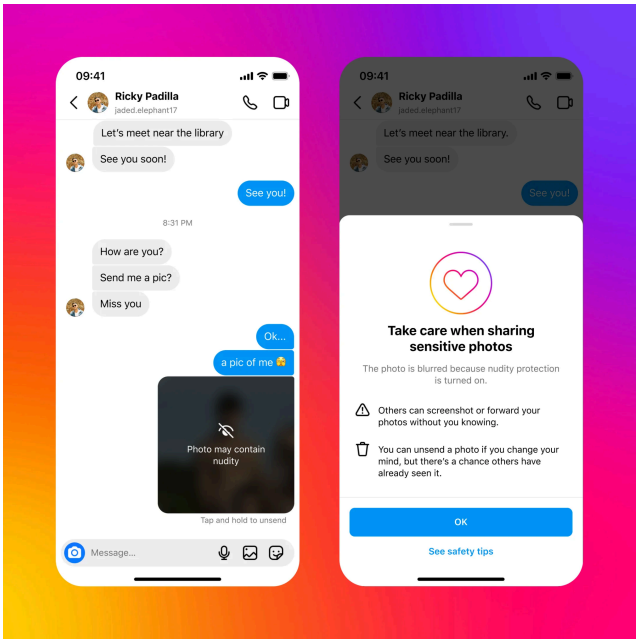
Outcome 2: Provide safeguards to reduce the risk of harm arising from online bullying or harassment

Measure 6. Implement, enforce and/or maintain policies and processes that seek to reduce the risk to individuals (both minors and adults) or groups from being the target of online bullying or harassment.

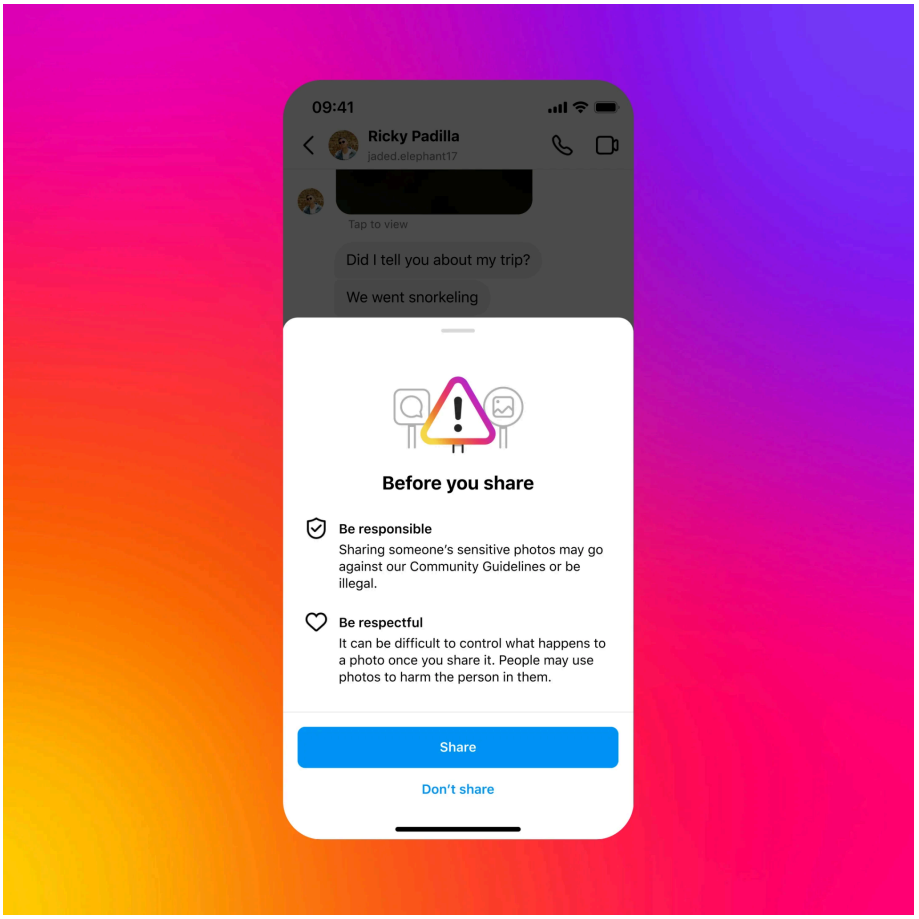
Measure 7. Implement and maintain products and/or tools that seek to mitigate the risk of individuals or groups from being the target of online bullying or harassment.

Measure 8. Implement, maintain and raise awareness of product or service related policies and tools for users to report online bullying or harassment content.

Measure 9. Support or maintain programs, initiatives or features that seek to educate and raise awareness on how to reduce or stop online bullying or harassment.
At Meta, we take bullying and harassment situations seriously. Recognising the complexity of these issues, which often require context-specific solutions, we strive to enforce against such content while also providing our community with tools to protect themselves in ways that are most effective for them. We continuously develop new resources and tools, and revise our policies with expert input, to ensure that we are effectively tackling this issue. Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to combat bullying and harassment, globally and in New Zealand, and have been published on the Code website. **Protecting young people from sextortion and intimate image abuse** • Our latest efforts include new tools we are testing to protect people from sextortion and make it more difficult for scammers to find potential targets on Meta's apps and across the internet. We are also testing new measures to support young people in recognising and protecting themselves from sextortion scams. • These updates build on our ongoing work to help protect young people from unwanted or potentially harmful contact. Safety Notices are shown to teens who are already in contact with potential scam accounts. We offer a dedicated option for people to report DMs that are threatening to share private images. We also supported NCMEC in developing Take It Down, a platform that lets young people take back control of their intimate images and helps prevent them being shared online, thereby taking power away from scammers. • In April 2024, we began testing a new feature in Instagram DMs that blurs images detected as containing nudity and prompts users to reconsider sending nude images. This feature aims to protect individuals from receiving unwanted nudity in their DMs, as well as from scammers who may use nude images to deceive people into sharing their own images. • By default, the nudity protection feature will be enabled for teenagers under the age of 18 globally. We will also display a notification to adults, encouraging them to opt-in to this feature. • When the feature is turned on, individuals attempting to send images containing nudity will receive a message cautioning them to be mindful when sharing sensitive photos and reminding them that they can unsend these images if they change their mind.

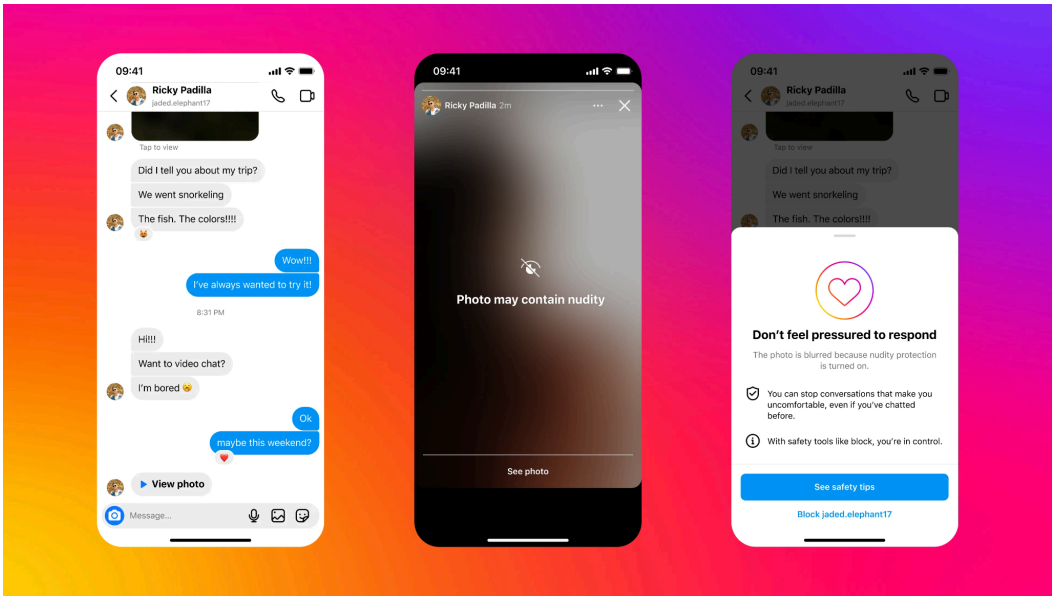


- In addition, anyone who tries to forward a nude image they have received will see a message encouraging them to reconsider.

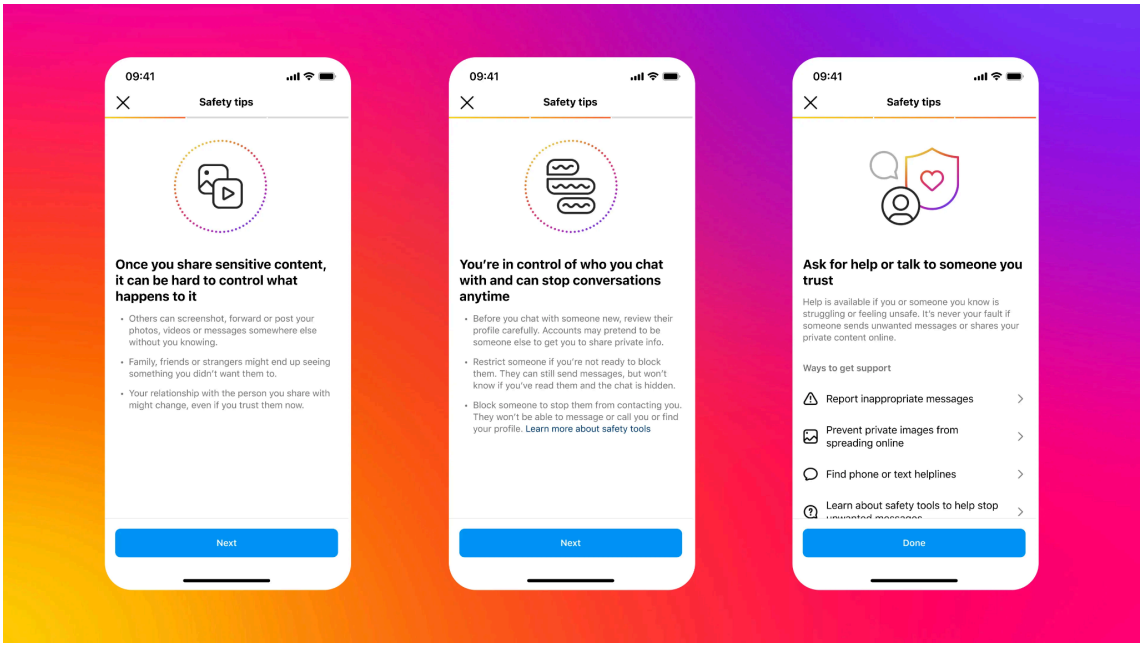


- When a user receives an image containing nudity, it will be automatically blurred and covered with a warning screen. This way, the recipient is not immediately exposed to the

nude image and has the option to choose whether or not to view it. Additionally, we will display a message to the recipient, urging them not to feel pressured to respond to the image. The message will also provide an option for the recipient to block the sender and report the chat if they wish to do so.



- When sending or receiving images containing nudity, users will be directed to safety tips developed with expert guidance. These tips highlight potential risks, such as the possibility of images being screenshotted or forwarded without consent, changes in relationships over time, and the importance of carefully reviewing profiles to ensure authenticity. They also link to a range of resources, including Meta's Safety Center, support helplines, StopNCII.org for those over 18, and Take It Down for those under 18.
- The nudity protection feature uses on-device machine learning to analyse whether an image sent in a DM on Instagram contains nudity. This analysis is performed directly on the device, ensuring that the feature remains functional even in end-to-end encrypted chats where Meta does not have access to the images, unless a user chooses to report them to us.



Meta partners with New Zealand organisations and experts to support community initiatives and research relating to combatting bullying or harassment. Some of our most recent efforts include:

- Partnered with Netsafe NZ and the Ministry of Education to launch six learning modules for young people and parents across covering online media literacy; Staying safe on Instagram; Owning your info on FB; Safety on Facebook and Instagram for parents; Staying safe in new digital spaces; and an Introduction for families to the metaverse.

Global metrics on bullying and harassment content that we took action on globally in 2023 and the proactive rate of content detected before people reported it.

Period 2023	Facebook	Instagram
Jan-Mar	6.9 million with proactive rate over 65%	6.6 million with proactive rate over 90%
Apr-Jun	7.9 million with proactive rate over 65%	6.8 million with proactive rate over 90%
Jul-Sep	8.3 million with proactive rate over 87%	8.4 million with proactive rate over 93%
Oct-Dec	7.7 million with proactive rate over 86%	8.8 million with proactive rate over 95%

For New Zealand, in 2023:

- **We took action on over 16,000 pieces of content on Facebook in New Zealand for violating our Bullying & Harassment policy.** Over 67% of this content was detected proactively before people reported it to us.
- **We took action on over 82,000 thousand pieces of content on Instagram in New Zealand for violating our Bullying & Harassment policy.** Over 96% of this content was detected proactively before people reported it to us.

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

Measure 10. Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of hate speech.

Measure 11. Implement and maintain products and tools that seek to prohibit or reduce the prevalence of hate speech.

Measure 12. Implement, maintain and raise awareness of product or service related policies and tools for users to report potential hate speech.

Measure 13. Support or maintain programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of online hate speech.

Measure 14. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online hate speech.

Meta does not tolerate hate speech on its platforms, including Facebook and Instagram. Hate speech is defined as violent or dehumanising speech, statements of inferiority, calls for exclusion or segregation based on protected characteristics such as race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disability or disease.

Meta has policies that prohibit hate speech on our services. 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. Sometimes, based on local nuance, we consider certain words or phrases as frequently used proxies for protected characteristic groups.

There may be instances where the definition of hate speech is unclear due to different meanings, intent, or context. As such, Meta provides an appeals process for users to dispute decisions made on enforced content.

Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to combat hate speech, globally and in New Zealand, and have been published on the Code website.

Our latest efforts include the following partnerships:

- Meta is a member of the Christchurch Call and contributes to its various working groups.
- In 2023, we continued our support for Sakinah Community Trust to run their annual Unity Week events across Christchurch, which seek to build social cohesion and unity, in particular targeting young people, working closely with local schools, sporting and other youth-focused organisations.

Global metrics on the pieces of hate speech content that we took action on globally in 2023 and the proactive rate of content detected before people reported it:

Period	Facebook	Instagram
Jan-Mar	10.7 million with proactive rate over 82%	5.1 million with proactive rate over 95%
Apr-Jun	18 million with proactive rate over 88%	9.8 million with proactive rate over 97%
Jul-Sep	9.6 million with proactive rate over 94%	7 million with proactive rate over 96%
Oct-Dec	7.4 million with proactive rate over 94%	7.4 million with proactive rate over 97%

For New Zealand, in 2023:

- **We took action on over 14,000 pieces of content on Facebook in New Zealand for violating our Hate Speech policy.** Over 69% of this content was detected proactively before people reported it to us.
- **We took action on over 44,000 pieces of content on Instagram in New Zealand for violating our Hate Speech policy.** Over 98% of this content was detected proactively before people reported it to us.

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

Measure 15. Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of content that potentially incites violence.

Measure 16. Implement and maintain products and tools that seek to prohibit or reduce the prevalence of content that potentially incites violence.

Measure 17. Implement, maintain and raise awareness of product or service related policies and tools for users to report content that potentially incites violence.

Measure 18. Support or maintain programs and initiatives that seek to educate users on how to

reduce or stop the spread of online content that incites violence.

Measure 19. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online content that incites violence.

Freedom of expression is a fundamental human right that underpins many other rights. However, we recognise that technologies designed to facilitate free expression, information, and opinion can also be exploited to disseminate hate speech and misinformation, which can incite violence. To address this challenge, we need to develop both short-term solutions that can be implemented during crises and a long-term strategy to ensure the safety of our users.

Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to combat violent extremism, globally and in New Zealand, and have been published on the Code website.

Our latest efforts include:

Updates to dealing with violent extremism

- **Dangerous Organisations and Individuals (DOI) Search Intercept:** In addition to our policies and technology, a key approach for Meta in dealing with violent extremist content is through various product interventions. One such intervention is the DOI Search Intercept, which is triggered when users search for terms related to DOI on platforms like Facebook and Instagram. This measure aims to restrict access to harmful content, guide at-risk users towards support services, and offer educational resources. This measure complements other efforts to reduce the visibility of content that breaches DOI policies.

Meta partners with New Zealand organisations and experts to support community initiatives and research relating to combatting incitement of violence. Some of our most recent efforts include:

- **Global Internet Forum to Counter Terrorism (GIFCT):** The GIFCT hash-sharing database allows participating companies to share hashes, or digital fingerprints, of known terrorist content. This process works by first creating a unique hash of the identified terrorist content on a company's platform. The company then shares this hash with other participating companies through the GIFCT database. When a user uploads content to a participating platform, the platform checks the uploaded content against the shared hashes in the database. If a match is found, the platform can automatically remove the content, helping to prevent the spread of terrorist material online.
- To help other platforms and entities that may not have the resources and the technology, we have developed and made available a free open source software tool called Hasher-Matcher-Actioner (HMA) that identifies copies of images or videos and takes action against them en masse. HMA builds on previous open source image and video matching software from Meta, and it can be used for any type of violating content. We hope the tool will be adopted by a range of companies to help them stop the spread of terrorist content on their platforms, and that it will be especially useful for smaller companies who don't have the same resources as bigger companies.

Global metrics on pieces of content that incite violence that we took action on globally in 2023 and the proactive rate of content we detected before people reported it:

Period	Facebook	Instagram
Jan-Mar	12.4 million with proactive rate over 97%	7.7 million with proactive rate over 98%
Apr-Jun	10.6 million with proactive rate over 97%	7.5 million with proactive rate over 98%
Jul-Sep	8.6 million with proactive rate over 85%	7.2 million with proactive rate over 98%
Oct-Dec	8.2 million with proactive rate over 86%	9.6 million with proactive rate over 99%

For New Zealand, in 2023:

- **We took action on over 48,000 pieces of content on Facebook in New Zealand for violating our Violence & Incitement policy.** Over 83% of this content was detected proactively before people reported it to us.
- **We took action over 54,000 pieces of content on Instagram in New Zealand for violating our Violence & Incitement policy.** Over 99% of this content was detected proactively before people reported it to us.

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

Measure 20. Implement, enforce and/or maintain policies and processes that seek to prohibit and/or reduce the spread of violent or graphic content online.

Measure 21. Implement and maintain products and tools that seek to and/or reduce the spread of violent or graphic content.

Measure 22. Implement, maintain and raise awareness of product or service related policies and tools for users to report potential violent and graphic content.

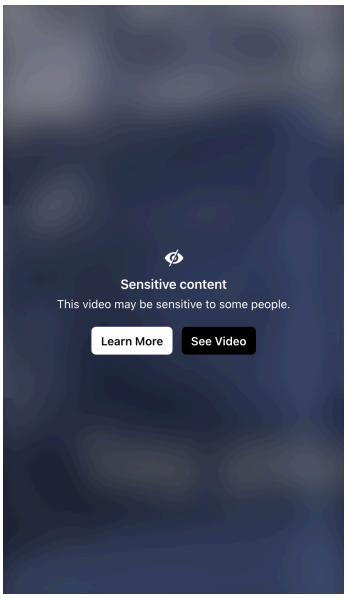
We remove content that glorifies violence or celebrates the suffering or humiliation of others on Facebook and Instagram. We do allow people to share some graphic content to raise awareness about current events and issues. In these cases, we may restrict the content from being viewed by people under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

Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to combat violent or graphic content, globally and in New Zealand, and have been published on the Code website.

Our latest efforts include:

Warning labels for sensitive content

- There are certain types of content that we may permit on our platform due to their public interest, newsworthiness, or value in promoting free expression. These categories of content may be disturbing or sensitive for some users, such as violent or graphic content that meets our exceptions list (e.g., it serves as evidence of human rights abuses or an act of terrorism).
- When a piece of content is identified as "disturbing" or "sensitive," we apply a warning label that restricts users from viewing the content unless they actively choose to do so. Additionally, the content will not be visible to users under the age of 18, and they will not be presented with the option to view it.



Global metrics pieces of violent and graphic content that we took action on globally in 2023 and the proactive rate of content detected before people reported it.

Period	Facebook	Instagram
Jan-Mar	13.6 million with proactive rate over 98%	5.1 million with proactive rate over 98%
Apr-Jun	13.8 million with proactive rate over 97%	6.2 million with proactive rate over 99%
Jul-Sep	9 million with proactive rate over 98%	4.1 million with proactive rate over 98%
Oct-Dec	14.6 million with proactive rate over 98%	17.5 million with proactive rate over 99%

For New Zealand, in 2023:

- **We took action on over 16,000 pieces of content on Facebook in New Zealand for violating our Violent and Graphic Content policy.** 97% of this content was detected proactively before people reported it to us.
- **We took action on over 25,000 pieces of content on Instagram in New Zealand for violating our Violent and Graphic Content policy.** 99% of this content was detected proactively before people reported it to us.

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

Measure 23. Implement, enforce and/or maintain policies, processes and/or products that seek to reduce the spread of online misinformation.

Measure 24. Implement, enforce and/or maintain policies and processes that seek to penalise users who repeatedly post or share misinformation that violates related policies.

Measure 25. Support or maintain media literacy programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of misinformation.

Measure 26. Support or maintain programs and/or initiatives that seek to support civil society, fact-checking bodies and/or other relevant organisations working to combat misinformation.

Measure 27. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from misinformation.

Misinformation is a complex social phenomenon which involves a range of offline and online behaviours. It goes beyond any single online platform — in fact, it long predates the internet. Misinformation can make the world less informed and erode trust. In some instances, it can even contribute to offline harm. All of us — tech companies, media companies, newsrooms, governments, civil society, educational institutions and citizens - have a role and responsibility to stop the spread of misinformation.

Our approach to misinformation is guided by our values of expression, safety, dignity, authenticity and privacy. 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.

Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to combat misinformation, globally and in New Zealand, and have been published on the Code website.

Our latest efforts include:

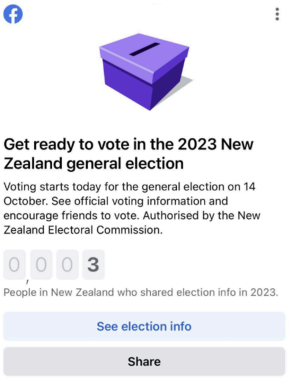
Combatting election-related misinformation

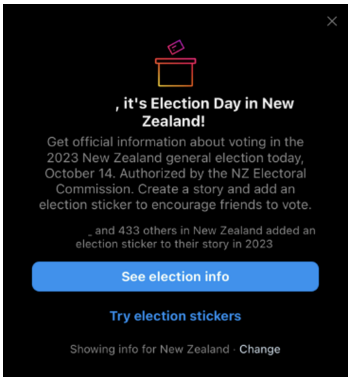
We remove content that has the potential to directly disrupt the functioning of political processes. This encompasses misinformation regarding voting and voter registration, such as false claims about online voting apps, as well as misinformation about voter eligibility, vote counting, and required materials for voting.

Voting is essential to democracy, which is why we take a firm approach on misrepresentations and misinformation that could result in voter fraud or interference.

Election day reminders:

- In the lead up to New Zealand's election in October 2023, Meta developed a comprehensive strategy[1] that focussed on proactively detecting and removing content that breached our policies, combatting misinformation and harmful content and promoting civic participation.
- Ahead of the October 2023 election in New Zealand, we developed election day reminders on Facebook and Instagram to encourage people to vote. Examples of these prompts for Facebook and Instagram, and stickers for Instagram can be found at the end of this case study.
- These prompts reached a large number of New Zealanders: 2.38m on Facebook & 1m on Instagram
- We also initiated the following specific programs of work in the lead up to the election, particularly in relation to combatting misinformation and disinformation. This work included:
    - **Combatting Misinformation and Expanding Capacity for Meta's Fact-Checkers in New Zealand:** To ensure that voters in New Zealand had access to reliable information ahead of the election, we provided a one-off funding boost to one of our fact-checking partners in NZ, Australian Associated Press, to increase their capacity in the lead up to the election. Our fact-checkers are independent and work to reduce the spread of misinformation across Meta's services. When they rate something as false, we significantly reduce its distribution so fewer people see it. We also notify people who try to share fact-checked content and add a warning label with a link to a debunking article.
    - **Empowering People to Identify False News:** Since we know it's not enough to just limit or remove harmful or misleading misinformation that people see, we launched a nation-wide media literacy campaign with Australian Associated Press (AAP). This shared tips and advice with people so that they could make informed decisions about what they read, trust and share. According to reporting shared by AAP, the six-week campaign, which was run on Facebook and Instagram, reached 2.48 million New Zealand adults and was viewed 13.88 million times.





[1] Meta on Medium (14 August 2023), 'How Meta is Preparing for the 2023 New Zealand Election'

Harmful health misinformation and imminent physical harm

- As the world was learning about the COVID-19 pandemic in January 2020, misleading and false information about COVID-19 was spreading on the internet and social media. By the end of that month, we had updated our policies to remove content with false claims or conspiracy theories flagged by leading global health organisations and local health authorities that could cause harm to people who believe them. We did it as an extension of our existing policies to remove content that could contribute to physical harm. We focused on claims designed to discourage treatment or taking appropriate precautions. The policy continued to evolve throughout the pandemic to address emerging false claims, including claims that could contribute to vaccine refusal.
- As the COVID-19 situation evolved, in mid-2022, we sought advice from Meta's Oversight Board specifically on whether we should change our approach to COVID-19-related misinformation. The Oversight Board issued an opinion finding that Meta should prepare measures for when the World Health Organisation lifted its public health emergency declaration, in order to protect freedom of expression and other human rights. In June 2023, we released our response to the Oversight Board's recommendations,[2] announcing that we would take a more tailored approach to our COVID-19 misinformation rules, in line with the Board's recommendations: in countries that still have a COVID-19 public health emergency declaration, we will continue to remove content for violating our COVID-19 misinformation policies, given the risk of imminent physical harm. Globally, our COVID-19 misinformation rules are no longer in effect, as the global public health emergency declaration that triggered those rules has been lifted.

Globally in 2023:

- We displayed warnings on over 543 million distinct pieces of content on Facebook (including reshares) globally based on over 238,000 debunking articles written by our fact checking partners.
- We displayed warnings on over 11 million distinct pieces of content on Instagram (including reshares) globally based on over 65,000 debunking articles written by our fact checking partners.

For New Zealand, in 2023:

- We displayed warning labels on over 2 million distinct pieces of content on Facebook in New Zealand (including reshares) based on over 62,000 articles written by our global third-party fact checking partners.
- We displayed warning labels on over 146,000 distinct pieces of content on Instagram in New Zealand (including reshares) based on over 17,000 articles written by our global third-party fact checking partners.

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

[2] Oversight Board, 'Oversight Board publishes policy advisory opinion on the removal of COVID-19 misinformation', April 2023.

Measure 28. Implement, enforce and/or maintain policies, processes and/or products that seek to suspend, remove, disable, or penalise the use of fake accounts that are misleading, deceptive and/or may cause harm.

Measure 29. Implement, enforce and/or maintain policies, processes and/or products that seek to remove accounts, (including profiles, pages, handles, channels, etc) that repeatedly spread disinformation.

Measure 30. Implement, enforce and/or maintain policies, processes and/or products that seek to provide information on public accounts (including profiles, pages, handles, channels, etc) that empower users to make informed decisions (e.g. date a public profile was created, date of changes to primary account information, number of followers).

Measure 31. Implement, enforce and/or maintain policies, processes and/or products that seek to provide transparency on paid political content (e.g. advertising or sponsored content) and give users more context and information (e.g. paid political or electoral ad labels or who paid for the ad).

Measure 32. Implement, enforce and/or maintain policies, processes and/or products that seek to disrupt advertising and/or reduce economic incentives for users who profit from disinformation.

Measure 33. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from disinformation.

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

To address disinformation, we employ a three-pronged approach: preventing interference, fighting misinformation, and increasing transparency. Our team of over 200 experts, with diverse backgrounds in law enforcement, national security, investigative journalism, cybersecurity, law, and engineering, works to disrupt networks of threat actors. We also continually improve our scaled solutions to detect and prevent the proliferation of inauthentic accounts and behaviours, and collaborate with civil society, researchers, and governments to strengthen our defences.

Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to combat disinformation, globally and in New Zealand, and have been published on the Code website.

Our latest efforts include:

Global adversarial threats

- In 2023, we took action against several Coordinated Inauthentic Behavior (CIB) networks.[3] More than half of these networks targeted audiences outside of their countries of operation. We were able to remove the majority of these networks before they could build a significant following. Our analysis of CIB trends in 2023 revealed four key patterns: an increase in CIB disruptions originating from China, the use of for-hire surveillance operations globally, abuse of domain name infrastructure, and the persistence of the Russian network 'Doppelgänger' in attempting to remain active online.

[3] Meta, Adversarial Threats report Q4 2023, Meta transparency Centre.

As reported in our quarterly Community Standards Enforcement Report, we have removed billions of fake accounts. The table below shows the number of accounts removed globally in 2023 and the proactive rate of fake accounts detected and actioned on before people reported them.

Period	Facebook	Instagram
Jan - Mar	426 million with proactive rate over 98%	not available
Apr - Jun	676 million with proactive rate over 98%	not available
Jul - Sep	827 million with proactive rate over 99%	not available
Oct - Dec	691 million with proactive rate over 99%	not available

4.2 Empower users to have more control and make informed choices

Signatories recognise that users have different needs, tolerances, and sensitivities that inform their experiences and interactions online. Content or behaviour that may be appropriate for some will not be appropriate for others, and a single baseline may not adequately satisfy or protect all users. Signatories will therefore empower users to have control and to make informed choices over the content they see and/or their experiences and interactions online. Signatories will also provide tools, programs, resources and/or services that will help users stay safe online.

Outcome 8. Users are empowered to make informed decisions about the content they see on the platform
Measure 34. Implement, enforce and/or maintain policies, processes, products and/or programs that helps users make more informed decisions on the content they see Measure 35. Implement, enforce and/or maintain policies, processes, products and/or programs that seek to promote accurate and credible information about highly significant issues of societal importance and of relevance to the digital platform's user community (e.g. public health, climate change, elections) Measure 36. Launch programs and/or initiatives that educate or raise awareness on disinformation, misinformation and other harms, such as via media/digital literacy campaigns
Outcome 9. Users are empowered with control over the content they see and/or their experiences and interactions online
Measure 37. Implement, enforce and/or maintain policies, processes, products and/or programs

that seek to provide users with appropriate control over the content they see, the character of their feed and/or their community online. Measure 38. Launch and maintain products that provide users with controls over the appropriateness of the ads they see.
A critically effective way to address online safety and harmful content is to build a resilient digital society by providing the tools and resources that will enable people to make informed decisions. By empowering individuals with the skills to critically evaluate information, we can foster a culture where people are critically effective in deciding what to read, trust, and share online. We do this by providing greater transparency and control to users; providing information that will help them make informed decisions; and advancing media and digital literacy. We offer many tools, products and resources to users to address different areas of safety risks and harms, including: • Authoritative information sources • Safety hubs • Warning labels and notices • Parental supervision and age-appropriate controls • Comments filtering tools • Context buttons with more information • Privacy tools • Controls to customise what users see in their Feed • Feed options that allows users to decide how they want content ranked Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to empower users to have more control and make informed choices, globally and in New Zealand, and have been published on the Code website. Our latest efforts include: **Updates to Reels and Feed ranking technologies** • In May 2024, we enhanced our Reels and Feed ranking technologies to more effectively deliver recommendations. We developed a new model architecture that can effectively learn from large datasets very efficiently, which led to significant improvements in our pilot with Facebook Reels. Over the next year or so, this advanced recommendations technology will be integrated into more products, including our entire video ecosystem and Feed recommendations. **Labelling AI-generated images** • As the distinction between human and synthetic content becomes increasingly ambiguous, people are seeking clarity on where the boundaries lie. With AI-generated content becoming

more prevalent, our users have expressed a desire for transparency regarding this emerging technology. Therefore, it is crucial that we assist individuals in identifying when photorealistic content has been created using AI.

- In February 2024, Meta announced its collaboration with industry partners to establish common technical standards for identifying AI content, including video and audio. In the coming months, we will label images posted by users on Facebook and Instagram when we detect industry-standard indicators that they are AI-generated. We have already labelled photorealistic images created using Meta AI since it launched, ensuring that users are aware that they are "Imagined with AI".

Updates to Facebook Feed AI system

- Each user's Facebook Feed is personalised based on their individual activity. Now, users have the ability to control or customise what they see. From May 2024, the content that a user sees on your Facebook Feed is selected, ranked and delivered to them by an artificial intelligence (AI) system. Within one AI system, multiple machine learning models work together to deliver a personalised experience. These models and their input signals are constantly evolving as the system learns and improves over time.
    - **Hide:** Users can hide a post so that they won't see that post again. This action also helps to minimise similar content from appearing in the Feed.
    - **Reconnect:** Users can reconnect to follow a person, Page or group that they unfollowed in the past.
    - **Snooze:** Users can temporarily stop seeing posts from a person, Page or group. They can also restart, stop or add more time to their snoozes.
    - **Report content:** If a user sees something they think goes against Meta's Community Standards, they can report it. This also applied to posts that they think are spam or false news.
    - **Show more/show less:** This feature lets users customise what you see in their Feed. Clicking "Show more" or "Show less" on a post will temporarily increase or decrease the ranking score for a post and other posts like it.
    - **Unfollow:** To stop seeing certain posts, users can unfollow a person, Page or group.
    - **Manage favourites:** Users can select people and Pages that they want to prioritise. This means that their posts will be shown higher in the Feed and users will see their newest posts first.
    - **See newest content first:** To view content from people, Pages and groups that users follow in reverse chronological order, visit the Feeds tab.
- To enhance Feed personalisation and boost user engagement, we leverage the AI system to make predictions about content that users will find most relevant and valuable. These prediction models utilise underlying input signals to select content that users are most likely to interact with. Below are some of the key predictions – and input signals that inform them – that we use in this AI system:

- The likelihood of users scrolling past a post without engaging with it
- The probability of users completing a video watch
- The likelihood of users clicking "Show more" if presented with the option below a post
- Predicting the number of times users will view the comments section within 1-24 hours after a post receives a comment
- Predicting if users' video play percentage is below a threshold, indicating they may not enjoy watching that video
- Estimating the amount of time users will spend viewing a post
- Assessing the likelihood of users engaging with a Page
- Predicting the likelihood of users tapping a story in feed
- Evaluating the probability of users engaging with a group or a post from that group
- Predicting the likelihood of users viewing additional comments on a post

4.3 Enhance transparency of policies, processes and systems

Transparency helps build trust and facilitates accountability. Signatories will provide transparency of their policies, processes and systems for online safety and content moderation and their effectiveness to mitigate risks to users. Signatories, however, recognise that there is a need to balance public transparency of measures taken under the Code with risks that may outweigh the benefit of transparency, such as protecting people's privacy, protecting trade secrets and not providing threat actors with information that may expose how they may circumvent or bypass enforcement protocols or systems.

Outcome 10. Transparency of policies, systems, processes and programs that aim to reduce the risk of online harms
Measure 39. Publish and make accessible for users Signatories' safety and harms-related policies and terms of service. Measure 40. Publish and make accessible information (such as via blog posts, press releases and/or media articles) on relevant policies, processes, and products that aim to reduce the spread and prevalence of harmful content online.
Outcome 11. Publication of regular transparency reports on efforts to reduce the spread and prevalence of harmful content and related KPIs/metrics
Measure 41. Publish periodic transparency reports with KPIs/metrics showing actions taken based on policies, processes and products to reduce the spread or prevalence of harmful content (e.g.

periodic transparency reports on removal of policy-violating content). Measure 42. Submit to the Administrator an annual compliance report, as required in section 5.4, that set out the measures in place and progress made in relation to Signatories' commitments under the Code.
Meta is committed to making transparent our safety and integrity-related policies, processes and systems where it does not pose a safety and security risk. We believe transparency promotes accountability by making platforms' efforts subject to public scrutiny and, in turn, holds us to account for the decisions we make. In our Baseline report, we laid out our general views and approach on transparency in section 2, and we have detailed our policies, processes (enforcement), tools and products (systems) in relation to the seven safety and harms themes in section 3. Information on our policies, processes and systems can be found in either our Transparency Center, Help Centers (Facebook, Instagram) or Newsroom. To date, Meta has published three transparency reports under the Code of Practice for Online Safety and Harms, with the latest launched in October 2024. Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to enhance transparency of our policies, processes and systems, globally and in New Zealand, and have been published on the Code website. Our 2024 report outlines the steps we took during the reporting period and over the 2023 calendar year to meet the 45 measures and 13 outcomes. We have taken steps to encourage accountability and oversight of our content decisions. Over five years ago, we proactively and voluntarily established an Oversight Board to make binding rulings on difficult and significant decisions about content on Facebook, Instagram and Threads. • **Oversight Board:** The Oversight Board also publishes quarterly Transparency Reports which provide new details on the Oversight Board's cases, decisions and recommendations. These quarterly updates are designed to 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. They are available in the dedicated Oversight Board Transparency Centre which is regularly updated to have the latest information on the Oversight Board's cases, recommendations, and appeals process. Between January and December 2023, the Oversight Board issued 53 decisions and 66 recommendations.[4] In August 2023, we launched the Meta Content Library and Content Library API. • **Meta Content Library and Content Library API:** These tools provide comprehensive access to the full public content archive from Facebook and Instagram. In 2024, we will make these research tools available to third-party fact-checking partners and qualified users globally, including New Zealand. Meta has partnered with the Inter-university Consortium for Political and Social Research (ICPSR) to provide researchers with access to public data from Meta's platforms. This collaboration enables researchers to apply for access to these

[4] Oversight Board, 'H2 2023 Transparency Report' and Oversight Board, 'Biannual and Quarterly Updates'.

valuable resources through ICPSR, while ensuring responsible and privacy-preserving data sharing practices.
- This is in addition to the **Meta Ad Library:** A searchable archive of all social issues and political ads on our services, we have progressively added functionality and real-time data on these ads.

In December 2023, Meta announced that we had commenced rolling out default end-to-end encryption for personal messages and calls on Messenger and Facebook, to help make these safer, more secure and private services.

- Encryption provides an extra layer of security which means that the content of users' messages and calls with friends and family are protected from the moment they leave the sender's device, to the moment they reach the receiver's device. It means that nobody, including Meta, can see what has been sent or said, unless the user chooses to report a message to us.[5]

4.4 Support independent research and evaluation

Independent local, regional or global research by academics and other experts to understand the impact of safety interventions and harmful content on society, as well as research on new content moderation and other technologies that may enhance safety and reduce harmful content online, are important for continuous improvement of safeguarding the digital ecosystem. Signatories will seek to support or participate in these research efforts.

Signatories may also seek to support independent evaluation of the systems, policies and processes they have implemented under the commitments of the Code. This may include broader initiatives undertaken at the regional or global level, such as independent evaluations of Signatories' systems.

Outcome 12. Independent research to understand the impact of safety interventions and harmful content on society and/or research on new technologies to enhance safety or reduce harmful content online.
Measure 43. Support or participate, where appropriate, in programs and initiatives undertaken by researchers, civil society and other relevant organisations (such as fact-checking bodies). This may include broader regional or global research initiatives undertaken by the Signatory which may also benefit Aotearoa New Zealand. Measure 44. Support or convene at least one event per year to foster multi-stakeholder dialogue, particularly with the research community, regarding one of the key themes of online safety and harmful content, as outlined in section 4. This may include broader regional or global events undertaken by the Signatory which involve Aotearoa New Zealand.
Outcome 13. Support independent evaluation of the systems, policies and processes that have been implemented in relation to the Code.

[5] Meta Newsroom, Launching Default End-to-End Encryption on Messenger

Measure 45. Commit to selecting an independent third-party organisation to review the annual compliance reports submitted by Signatories, and evaluate the level of progress made against the Commitments, Outcomes and Measures, as outlined in section 4, as well as commitments made by Signatories in their Participation Form (see Appendix 2).
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. Previous reports (2022 Baseline Report, 2023) have outlined the measures taken by Meta to support independent research and evaluation, globally and in New Zealand, and have been published on the Code website. Some of our most recent efforts include: ● **Providing seed funding for a collaboration to address online safety in the Pacific:** In 2023, Meta provided seed funding to establish an online safety collaboration between Save the Children Australia, ChildFund Australia and Netsafe New Zealand. The collaboration enables these organisations to provide additional education, services, and support to combat online bullying and harassment through a pilot initiative in Papua New Guinea. ● **Netsafe's Building Bridges Conference in Christchurch:** In 2023, Meta sponsored the Breaking Glass Ceilings and Building Bridges Conference, a trans-Tasman tech conference in Christchurch. Sessions covered how tech was used to combat violent extremism, sextortion, and include further representation of Māori, Pacific Nations, and LGBTQIA+ perspectives. Meta also participated in a panel session on online safety regulation. ● **AI Alliance:** Together with IBM in December 2023, Meta established the AI Alliance, a community of over 50 technology creators, developers and adopters collaborating to advance safe, responsible AI rooted in open innovation. The Alliance seeks to (i) build and support open technologies across software, models and tools, (ii) enable developers and scientists to understand, experiment, and adopt open technologies, and (iii) advocate for open innovation with organisational and societal leaders, policy and regulatory bodies, and the public. ● **Foundational Integrity Research:** In September 2022, Meta launched the Foundational Integrity Research awards to collaborate with academic institutions and researchers on foundational and applied research in areas such as misinformation, hate speech, violence and incitement, and coordinated harm. This initiative aims to foster deep collaborations and drive meaningful research in these critical domains, and received 500 proposals from over 300 universities and institutions globally.