Meta

Meta's Submission to the Joint Select Committee on Social Media and Australian Society

JUNE 2024

Executive Summary

Meta welcomes the opportunity to provide the Joint Select Committee on Social Media and Australian Society with this submission to assist it in its inquiry on the role and impact of social media in Australia.

Australians use many different apps to connect with their friends and family and with organisations, communities and local issues.[1] When Australians are using Meta's family of apps, we recognise that we have a responsibility to keep people safe, to comply with all applicable laws, to be responsive to community concerns, and to promote accountability and transparency.

Keeping people safe online has been a challenge since the start of the internet. As threats and trends constantly evolve, it's important that we continue to adapt so that people have safe and positive experiences across Meta's services. We are focused on building technology that people find useful and feel safe when doing so.

That is why we have around 40,000 people overall working on safety and security, and we have invested over US$20 billion since 2016. This includes around US$5 billion in the last year alone. This investment includes building and maintaining our content governance and integrity systems, as well as user transparency tools and controls, and partnerships and programs through which we receive feedback and promote digital skills and literacy.

With respect to content governance, we use a strategy called "remove, reduce, inform"[2] to manage content across Meta technologies. This means that we remove harmful content that goes against our policies,[3] reduce the distribution of problematic content that doesn't violate our policies, and inform people with additional context so they can decide what to click, read or share. We also offer a range of tools so that people can customise their experience above and beyond the baseline investment we make in safety and security.

To help with this strategy, we have policies that describe what is and isn't allowed on our technologies. Our teams work together to develop[4] our policies and enforce[5] them. Increasingly, we have been deploying proactive detection technology to identify and action harmful content before anyone reports it to us. For many categories, our proactive rate (the percentage of

[1] On average, Australians use 6.1 different social platforms each month (DataReportal Digital 2024: Australia report, slide 56, https://datareportal.com/reports/digital-2024-australia)
[2] Meta, 'Taking down violating content', *Transparency Center*, https://transparency.meta.com/en-gb/enforcement/taking-action/taking-down-violating-content/
[3] Meta, 'Policies', *Transparency Center*, https://transparency.meta.com/en-gb/policies/
[4] 'How stakeholder engagement helps us develop the Facebook Community Standards', *Transparency Center*, https://transparency.meta.com/en-gb/policies/improving/stakeholders-help-us-develop-community-standards/
[5] Meta, 'How Meta prioritises content for review', *Transparency Center*, https://transparency.meta.com/en-gb/policies/improving/prioritizing-content-review/

content we took action on that we found before a user reported it to us), is more than 99 per cent across high-risk content types.

Beyond actioning harmful content, we also work to promote a safe and positive experience on our services by using technology and offering tools to help users customise their use of Meta's services. These features are informed by our consultations with industry, experts, and civil society organisations, including in Australia.

In particular, we are working to give teens more age-appropriate experiences. We place teens into the most restrictive content recommendation settings on Instagram and Facebook[6], and place teens into stricter messaging and reachability default settings so that they are not able to receive messages from anyone they are not connected to and/or potentially suspicious accounts[7].

We recognise the importance of being responsive to the community. This is why we regularly undertake surveys of our community, work closely with experts in the development of our policies, and develop long-standing partnership for feedback and joint programmatic work to promote awareness of our policies, tools and tips for having a positive experience online. For example, in November 2023, we partnered with the Australian Federal Police-led Australian Centre to Counter Child Exploitation, Kids Helpline and US-based organisation NoFiltr (Thorn) to inform young people about sextortion. We also have a local Online Safety Advisory Group and a local Combatting Hate Advisory Group.

To help inform the public debate about our content governance and integrity systems, we publish a range of transparency reports, many quarterly. These include the Community Standards Enforcement Report,[8] the Adversarial Threat Report[9] and the Government Requests for User Data Report.[10] Since 2017, we've reported takedowns of more than 200 covert influence operations, cyber espionage, mass reporting and brigading networks.

Our policies and transparency measures are not just focused on the content that we remove and action. We also have policies and transparency tools that provide details about the policies that govern how content is distributed and recommended on our services, more than 22 publicly available Systems Cards explaining recommender systems for content across Facebook and

[6] Meta, 'New Protections to Give Teens More Age-Appropriate Experiences on Our Apps', *Newsroom*, 9 January 2024, https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps
[7] Meta, 'Introducing Stricter Message Settings for Teens on Instagram and Facebook', *Newsroom*, 25 January 2024, https://about.fb.com/news/2024/01/introducing-stricter-message-settings-for-teens-on-instagram-and-facebook
[8] Meta, 'Community Standards Enforcement Report', *Transparency Center*, https://transparency.meta.com/reports/community-standards-enforcement/
[9] Meta, 'Meta's threat disruptions', *Transparency Center*, https://transparency.meta.com/en-gb/metasecurity/threat-reporting/
[10] Meta, 'Government Requests for User Data', *Transparency Center,* https://transparency.meta.com/reports/government-data-requests/

Instagram,[11] as well as on-platform tools that let people understand why they are seeing a particular piece of content or ad and the ability to change it.[12]

Our work does not just stop at Meta's family of apps – we have also been working to promote standard setting, deeper collaboration and accountability more broadly across the industry.

Given the diversity of apps that Australians use to communicate with and given the increased adversarial nature of the online ecosystem, we understand that people expect to have the same protections across a wide range of apps. This is why Meta has been at the forefront of working both globally and within Australia on industry codes and cross-industry collaborations to promote accountability, transparency and research and investment to promote safety and security.

Globally, we were a founding member of the Global Internet Forum to Counter Terrorism[13] and we are also a member of the Tech Coalition, a global alliance of technology companies that work together to drive critical advances in technology and adoption of best practices for keeping children safe online.[14] As part of the Coalition, we have been a founding member of the Lantern program, which enables technology companies to share signals about accounts and behaviors that violate their child safety policies. We provided the Tech Coalition with the technical infrastructure that sits behind the program as well as oversee the technology with them, ensuring it is simple to use and provides our partners with the information they need to track down potential predators on their own platforms.

In Australia, we are a founding member of the DIGI Code of Practice on Disinformation and Misinformation, recently publishing our fourth transparency report under that Code. We have also contributed to development of the six industry codes that were recently registered by the eSafety Commissioner under the Online Safety Act, and have been driving the development of a scams code that will shortly be released by DIGI.

And finally, we recognise the importance of greater transparency, accountability and user empowerment that is required for smart regulatory frameworks. This is why we have long been calling[15] for government regulation for digital platforms, working to establish proactive regulatory models (such as the codes mentioned above) and contributing constructively to the debate surrounding digital policy. Given the many new laws that have been enacted or proposed in Australia specifically focused on digital platforms, the question is no longer whether regulation is needed, but whether it is effective at driving appropriate investment in safety and

[11] Meta, 'Building Generative AI Features Responsibly', *Newsroom*, 27 September 27 2023, https://about.fb.com/news/2023/09/building-generative-ai-features-responsibly/
[12] Meta, 'Our approach to explaining ranking', *Transparency Center*, https://transparency.meta.com/features/explaining-ranking/
[13] Global Internet Forum to Counter Terrorism, https://gifct.org/
[14] Tech Coalition, https://www.technologycoalition.org/what-we-do
[15] Meta, 'Four Ideas to Regulate the Internet', 30 March 2019, https://about.fb.com/news/2019/03/four-ideas-regulate-internet/

security across the industry. We welcome the opportunity to contribute to the debate around the efficacy of these regulatory frameworks.

The efficacy of digital platform regulation will inevitably be challenged as technology and the way people use it changes over time. The Facebook app began as a primarily text based experience on desktop; when Instagram was first released it was solely for sharing photos – now 50% of the time that is spent within the app occurs with Reels. As consumer preferences have changed and there has been an increase in the number of apps and services that people can use to connect and to entertain and inform themselves, Meta has invested significantly to compete with other services and to innovate with the goal of continuing to provide the people who use our services with useful and enjoyable experiences on our services, evolving them over time. From time to time, however, the new product innovations that we have built have not met consumer needs and trends and have been withdrawn from the market. This has been the case for Facebook News, which we announced we would be deprecating in Australia in February 2024, consistent with the deprecation of this product in the US, UK, France and Germany. This also means that the deals that underpin this product will not be renewed. The deprecation of Facebook News does not, however, change our commitment to or our investment in our content governance and integrity systems, including those to combat disinformation and misinformation.

We welcome the opportunity to assist this Committee with its inquiry as part of our ongoing commitment to ensure that our services play a positive role in Australia.

Table of Contents

Impact of Meta's services in Australia

Millions of Australians regularly use Meta's family of apps. People use our apps to connect with family and friends, join a community group focused on an interest or passion, to connect with public figures, small businesses and other organisations. For example, more than 1.8 billion people engage in Facebook Groups every month.[16]

One example of this occurred during the Women's World Cup last year when The Matildas and A-League Women Supporters Facebook Group mobilised their members, and organised group meet-ups at games around Australia and New Zealand for their 32,000 members. The Group provided a support network for women online and in real life for these fans, helping them to organise everything from securing tickets to games, providing safe spaces to fans to meet-up and form new bonds over their shared love for the sport which swept Australia last year.

In addition to individual Australians connecting with people and the communities they care about, digital platforms such as Facebook and Instagram have contributed to a democratisation of ecommerce, allowing anyone with an idea to test it out online and then use free and low-cost digital advertising products to grow their business. In relation to small businesses in Australia, a Deloitte report found that 82% of Australian small businesses reported using free, ad-supported Facebook apps to help them start their business, and 71% of Australian small businesses that use personalised advertising reported that it is important for the success of their business.[17]

For example, health business SaltLab makes magnesium salt spray in Melbourne and uses Facebook and Instagram to find new customers all around Australia and the world. Between 2022 and 2033, they experienced x3 the growth because they found more and more customers online using our platforms and as a result, they had to triple the amount of warehouse space and staff to keep up with the demand.

Another example is SafeStyle, who has said: "We launched SafeStyle on Instagram in 2018. We saw a gap in the market and thanks to Meta's personalised ads products, SafeStyle has been able to tailor its marketing to our core audience of tradesmen, miners, healthcare workers & weekenders who are looking for quality, comfortable and stylish safety glasses to protect them 'from the worksite to the weekend'. As a result over the last four years SafeStyle's online sales have grown from $1,000 to $1,000,000 a month, and we've grown our team from 1 to 30 people."[18]

---

[16] Meta, 'The Future of Facebook', *Newsroom*, 31 May 2024, https://about.fb.com/news/2024/05/the-future-of-facebook/
[17] Deloitte, *Dynamic Markets Unlocking small business innovation and growth through the rise of the personalized economy,* May 2021, https://scontent-syd2-1.xx.fbcdn.net/v/t39.8562-6/10000000_4303078769743544_7237603050373993547_n.pdf?_nc_cat=109&ccb=1-7&_nc_sid=e280be&_nc_ohc=Ondy-BSRxlsQ7kNvqFscXjm&_nc_ht=scontent-syd2-1.xx&oh=00_AYDuKsAxzsK1phNKn4yTqcC8kgyh0S_s7Yr8XkgNWhZulg&oe=6682D189
[18] Meta, *Digital Transformation: find your audience, build your brand*

To better understand the experience of Australian small business use of digital platforms such as Facebook and Instagram, in November 2023, we partnered with the Council of Small Business Organisations of Australia (COSBOA) to launch new research, 'The Digital Journey of SMEs in Australia.'[19] The research was conducted by Thoughtlab, which estimated in the report that digital technology allowed small and medium enterprises in Australia to generate $306 billion in additional revenue over the last year. If all SMEs in Australia became digital leaders, that could potentially unlock a further $181 billion in revenue. Other key findings from the report included:

- 79% of surveyed SMEs agreed that digital technologies are important or very important to them for driving innovation.
- 78% of digital leaders say digital technology helped them to increase business revenue over the past 12 months.
- 77% of SMEs reported that Meta's platforms helped people learn about their business.
- 76% reported that Meta's platforms help the business build customer relationships.
- 67% of SMEs believed their business was stronger because of Meta technologies and apps and 61% say their performance would suffer if they lost access to Meta technologies.

To invest in the digital skills of Australian small businesses, the Meta Boost Australia initiative offers free digital skills workshops that provide small businesses with the tools they need to start and grow a business online. Since the launch of Boost in 2018, Meta has held workshops in over 50 locations and trained over 30,000 Australian small businesses. Most recently, we have delivered Meta Boost sessions in Western Sydney (May 2023), Byron Bay (November 2022), Dubbo (March 2022) and the Blue Mountains (May 2021). In 2024, we will further our partnership with COSBOA to deliver new Boost training in Australia, delivering sessions on optimising the use of short-form video, Generative AI and digital advertising, as well as a module on cybersecurity and how to identify and avoid scams.

Another trend that has arisen with the growth is usage of digital platforms is the rise of the 'creator economy'[20]. In 2020, the overall size of the Creator Economy was estimated to be more than $100 billion.[21] A 2022 survey carried out by Edelman Data & Intelligence for Adobe found that across nine large nations, including Australia, there are more than 300 million Creators. In Australia, 6 million people self-identify as Creators.[22]

Creators express themselves and connect with their communities using multiple formats including images and video via Feed, Stories, Live, and Reels across Instagram and Facebook,

[19] Thoughtlab, *The Digital Journey of SMEs in Australia,* https://thoughtlabgroup.com/the-digital-journey-of-smes-in-australia/
[20] R Florida, *The Rise of the Creator Economy*, November, 2022, https://creativeclass.com/reports/The_Rise_of_the_Creator_Economy.pdf
[21] NeoReach and Influencer Marketing Hub, *Creator Earnings Breakdown: Where Are We in the Creator Economy*, 26 May 2021, https://influencermarketinghub.com/ebooks/Creator_Economy_-_Creator%20Earnings_Benchmark_2021.pdf.
[22] Adobe, *Creators in the Creator Economy: A Global Study*, 25 August 2022, https://s23.q4cdn.com/979560357/files/Adobe-'Future-of-Creativity'-Study_Creators-in-the-Creator-Economy.pdf?trk=article-ssr-frontend-pulse_little-text-block. The survey was conducted by Edelman Data & Intelligence.

and more recently, Threads. Creators also come to our platforms to partner with brands, sell their own merchandise, and earn money from their supporters and community.

To support Australian creators, we have undertaken a range of initiatives, including:

- launching the Creator Academy to help support sustainable creators in Australia. The Academy supported aspiring and emerging creators as they navigate their journeys and careers by leveling up their understanding of Reels, monetisation and Branded Content, and other innovative tools,
- developing Instagram University, an educational event to support social publishers, creators and social media managers as they grow, monetise and build their respective careers, profiles and brands,
- launching Creator Lab, a virtual creator educational series for all Australians to attend, particularly those outside of major capital cities. Content was delivered by successful established creators for emerging and aspiring creators, to support them as they start to navigate their careers,
- partnering with Screen Australia for the third year in a row to deliver mentoring and content funding for First Nations creators on Instagram under the First Nations Creator Program,
- partnering with Sony Kando to educate and support 60 emerging art and photography creators,
- holding masterclass workshops for aspiring comedy, food, and fashion creators for Instagram and Threads, and
- launching *Reels Squad,* where we received over 300 applications from Australian creators, to come together, create content and build community over their shared creative journeys.

And finally, we also know that people in Australia use our services to connect with loved ones, request and offer support, and seek up-to-date information, warnings and alerts during a disaster. This is why in July 2023, Meta developed and ran a new training series called ‘Connect, Alert, Inform’, for emergency response organisations across Australia, New Zealand and the Pacific Islands. The training focused on helping to strengthen emergency communicators’ skills in using Meta’s services to build community and deliver critical disaster-related information.

The virtual training curriculum was developed in consultation with a disaster communications academic from the University of Technology Sydney, and in partnership with Emergency Management and Public Affairs (EMPA) and RMIT CrossCheck. It comprised sessions specifically tailored for emergency responders across a range of topics, including Meta’s crisis management tools; planning social media content; page moderation and account security; combating disaster related mis and dis-information; and social media advertising.

Content Governance & Integrity

Content Governance Framework

Policies & enforcement

We are committed to giving people a safe and positive experience when they use Meta's services - especially young people. It is essential to our business: people in Australia and around the world will only continue to use our services if they feel welcome and safe in doing so. Meta makes significant, industry-leading investments to protect the safety of the community on our platform.

Policies

Our policies, known as our Facebook Community Standards and Instagram Community Guidelines,[23] outline what is and is not allowed on Facebook and Instagram. These policies are developed based on a range of values to help combat abuse. Safety is a core value of our policies, alongside privacy, authenticity, voice, and dignity.[24]

Our policies prohibit various categories of harmful content, including child exploitation, adult sexual exploitation, violent and objectionable content, suicide and self-injury including eating disorders, bullying and harassment, hate speech and privacy violations.

We have developed these policies based on feedback from our community and the advice of experts in fields such as technology, public safety, child safety and human rights. To ensure that everyone's voice is valued, we take great care to craft policies that are inclusive of different views and beliefs, in particular those of people and communities that might otherwise be overlooked or marginalised.

Meta's policies are also regularly updated to keep pace with changes happening online and offline around the world. We regularly host a Policy Forum meeting to discuss potential changes to our policies and their enforcement. A variety of internal and external subject matter experts participate in this meeting and hear input from external groups. In keeping with our commitment to greater transparency, the minutes of these meetings are made publicly available.[25] A change log of changes made to each policy area is available within the Community Standards.[26]

---

[23] See Meta, *Community Standards*, https://www.facebook.com/communitystandards
[24] Monika Bickert, 'Updating the values that inform our community standards', 12 September 2019, https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/
[25] Meta, 'Policy Forum Minutes', https://transparency.meta.com/en-gb/policies/improving/policy-forum-minutes/
[26] See, for example, Meta, 'Facebook Community Standards - Child sexual exploitation, abuse and nudity', *Transparency Center*, https://transparency.meta.com/en-gb/policies/community-standards/child-sexual-exploitation-abuse-nudity/

Enforcement

In order to enforce our policies, we invest significantly in both technology and people to help detect violating content and suspicious behaviour.

We have built up teams of experts who work in this space and have around 40,000 people dedicated to keeping people safe on our apps.

We encourage users to report content that they are concerned about. Once reported, we assess these reports and take action on the content consistent with our policies. We have also been investing in proactive detection technology to identify and action harmful content before anyone sees it and reports it to us.

We have scaled our enforcement to review millions of pieces of content across the world every day, and use our technology to help detect and prioritise content that needs review. We continue to build technologies like RIO,[27] WPIE[28] and XLM-R[29] that can help us identify harmful content faster, across languages and content type (i.e. text, image, etc.). These technologies alongside our continued focus on AI technologies help us to scale our efforts quickly in keeping our platforms safe.

To provide transparency to the community that can be used to hold us to account, we provide data about our enforcement work in our Community Standards Enforcement Report.[30] The report is released quarterly and includes metrics such as how much content we are actioning, and what percentage was detected proactively. Currently, we report these metrics against14 policy areas on Facebook and 12 on Instagram.

The Community Standards Enforcement Report demonstrates the progress we have made in detecting and actioning content that violates our policies. Today, most of the content that violates our policies is detected by machine learning tools and is actioned before users report it to us. For many categories, our proactive rate (the percentage of content we took action on that we found before a user reported it to us), is well over 90 per cent across high-risk content types such as child exploitation material and terrorist content, as well as violent and graphic content.

---

[27] Reinforcement Integrity Optimiser (RIO). RIO is an end-to-end optimised reinforcement learning (RL) framework. It's used to optimise hate speech classifiers that automatically review all content uploaded to Facebook and Instagram. For more information visit https://ai.facebook.com/blog/training-ai-to-detect-hate-speech-in-the-real-world/

[28] Whole Post Integrity Embeddings (WPIE) is a pretrained universal representation of content for integrity problems. WPIE works by trying to understand content across modalities, violation types, and even time. Our latest version is trained on more violations, and more training data overall. This approach prevents easy-to-classify examples from overwhelming the detector during training, along with gradient blending, which computes an optimal blend of modalities based on their overfitting behaviour. For more information visit https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech/

[29] XLM-R uses self-supervised training techniques to achieve state-of-the-art performance in cross-lingual understanding, a task in which a model is trained in one language and then used with other languages without additional training data. Our model improves upon previous multilingual approaches by incorporating more training data and languages. For more information visit https://ai.facebook.com/blog/-xlm-r-state-of-the-art-cross-lingual-understanding-through-self-supervision/

[30] Meta, 'Community Standards Enforcement Report', *Transparency Center*, https://transparency.fb.com/data/community-standards-enforcement/

Online safety partnerships

We have over 400 safety partners across the world, including a number of partnerships in Australia, to ensure that our global safety efforts are complemented by local expertise.

Globally, we collaborate across industry through organisations like the Tech Coalition, an industry association dedicated solely to eradicating child sexual exploitation and abuse online. Most recently in November 2023, we became a founding member of the Lantern program,[31] which enables technology companies to share signals about accounts and behaviors that violate their child safety policies. We provided the Tech Coalition with the technical infrastructure that sits behind the program as well as oversee the technology with them, ensuring it is simple to use and provides our partners with the information they need to track down potential predators on their own platforms. This builds on the work of Project Protect[32] – an industry effort launched in 2020 to combat online child sexual abuse.

We are also a part of the WePROTECT Global Alliance[33] industry committee. WeProtect brings together experts from government, the private sector and civil society to protect children from sexual exploitation and abuse online.

Globally, we have a Safety Advisory Council[34], which comprises leading safety organisations and experts from around the world. Council members provide expertise and perspective that inform Meta's approach to safety. The Australian youth anti-bullying organisation PROJECT ROCKIT is one of 11 organisations globally that serves on this Council.

In Australia, we invest significantly in local organisations to promote important safety and wellbeing messages. For example, we have invested in a Digital Ambassadors program delivered by PROJECT ROCKIT.[35] Digital Ambassadors is a youth-led, peer-based anti-bullying initiative. A Digital Ambassador aims to utilise strategies to safely connect and tackle online hate. This is a more than decade-long- partnership that has directly empowered more than 25,000 young Australians to tackle cyberbullying.[36]

We have also developed an Australian Online Safety Advisory Group to consult and provide a local perspective on policy development. This group comprises experts such as CyberSafety Solutions, PROJECT ROCKIT, WESNET, and ReachOut, as well as many others.

---

[31] Meta, 'Introducing Lantern: Protecting Children Online', *Newsroom*, 7 November 2023, https://about.fb.com/news/2023/11/lantern-program-protecting-children-online/
[32] Meta, 'Facebook Joins Industry Effort to Fight Child Exploitation Online'. *Newsroom*, 11 June 2020, https://about.fb.com/news/2020/06/fighting-child-exploitation-online/
[33] WeProtect Global Alliance, https://www.weprotect.org/
[34] Meta, 'Learn more about the Meta Safety Advisory Council', *Help Center*, https://www.facebook.com/help/222332597793306
[35] PROJECT ROCKIT, *Launching: Digital Ambassadors,* https://www.projectrockit.com.au/digitalambassadors/
[36] R Thomas, 'Young People at the Centre', Meta Australia Policy Blog, *Medium*, 8 February 2021 (updated 28 January 2023), https://medium.com/meta-australia-policy-blog/young-people-at-the-centre-25142d16c0cf

In addition, we provide significant support to our safety partners to ensure that our users - especially young people - can connect and communicate safely. Most recently, we have funded and supported the following online safety and mental health initiatives:

- **PROJECT ROCKIT**: In November 2023, we partnered with youth-driven organisation PROJECT ROCKIT to create 'Intimate Images Unwrapped', a series of educational videos that aimed to build greater literacy and awareness around the dynamics of sharing of intimate images.
- **ReachOut:** In 2023, we partnered with youth mental health service, ReachOut, to launch a creator-led campaign aimed at fostering social and emotional wellbeing in the lead-up to, and following, the Voice to Parliament referendum. The campaign focused on supporting and empowering young First Nations people in navigating the complex social and emotional wellbeing challenges resulting from the referendum and its surrounding debate.
- **Kids Helpline and ACCCE:** In November 2023, we partnered with the Australian Federal Police-led Australian Centre to Counter Child Exploitation, Kids Helpline and US-based organisation NoFiltr (Thorn) to inform young people about sextortion. The campaign included educational resources encouraging preventative behaviours online, the signs to look out for, where to report and where to seek support.
- **Butterfly Foundation:** In May 2024, we launched 'Enter the Chat', an education campaign that brought together a group of Australian creators to discuss the impact that certain types of online content may have on body image, how to create content more consciously and what safety tools are available on Instagram to support body image and wellbeing.

General safety tools

In addition to our global investment in policies and their enforcement, we also develop tools that allow people to customise their experience above and beyond our content governance and integrity systems. We recognise that we cannot know each person's individual situation and these tools allow people to adjust the settings to suit their specific situation.

In addition to the long-standing tools of Block, Report, Hide, Unfollow,[37] we continue to introduce new features to help users manage their experience. These tools are informed by our consultations with industry, experts and civil society organisations.

Some of our more recent tools include:

- **Restrict:** on Instagram, if a user restricts someone, the other person will not be able to see when the user is online or if they have read their messages. Their new comments on users' posts will only be visible to them. Users can choose to see the comment, and then approve, delete or ignore it. THey will not receive any notifications for future comments

[37] We provide an overview of these and other tools in the Meta Safety Center: https://about.meta.com/actions/safety

from that person.[38] This feature was developed in direct response to feedback from teens who told us that Blocking can be too severe and they wanted a way to protect themselves, but still be able to keep an eye on a bully's activity.

- **Hidden Words:** allows people to filter comments or messages requests that contain offensive words or emojis.[39]
- **Limits**: allows a person to temporarily limit contact from anyone not on their close friends list or recent followers.[40]

Combatting child sexual abuse material

We have strict policies against child sexual abuse material as well as the sexualisation of minors and other activities that can lead to child exploitation. We are constantly improving and developing new technologies to stay ahead of these harms and have made some of these available to other companies at no cost.

We use a combination of technology and behaviour signals to detect and prevent child sexual abuse material, including grooming or potentially inappropriate interactions between a minor and an adult. We have invested in cutting edge technology to detect and remove this content. For example, we have developed two technologies (called PDQ and TMK+PDQF)[41] to detect identical and near-identical photos and videos, and we have made these technologies available open source and free of charge to allow industry partners, small developers and NGOs to also benefit from this technology.

This work has had a significant impact; in the last quarter alone, we removed 14.4 million pieces of child endangerment content and 94.3% of this content was detected and removed by us proactively before a user saw and reported it to us.[42] We also automatically disable accounts if they exhibit a certain number of the 60+ signals we monitor for potentially suspicious behavior.

We also work closely with the National Center for Missing and Exploited Children (NCMEC), a nonprofit organisation that refers cases to law enforcement in Australia and around the world, in compliance with US law. When we become aware of this content on our services, we report it to NCMEC directly. In the first quarter of 2024, we made 5.2 million NCMEC Cybertip Reports[43] for child sexual exploitation. Of these, over 90,000 involved inappropriate interactions with

[38] Meta, 'Stay Safe', *Safety Center*, https://about.meta.com/actions/safety/topics/safety-basics/tools/stay-safe
[39] Instagram, 'Introducing new tools to protect our community from abuse', 21 April 2021, *Help Center*, https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse and Instagram, 'Updates to How We Protect Our Community from Abuse', *Help Center*, https://about.instagram.com/blog/announcements/creator-safety-tools
[40] Instagram, 'Temporarily limit people from interacting with you on Instagram', *Help Center*, https://help.instagram.com/4106887762741654
[41] https://about.fb.com/news/2019/08/open-source-photo-video-matching/
[42] Meta, 'Community Standards Enforcement Report - Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation', *Transparency Center*, https://transparency.meta.com/reports/community-standards-enforcement/child-nudity-and-sexual-exploitation/facebook/
[43] CyberTips relating to inappropriate interactions with children may include an adult soliciting child sexual abuse material (CSAM) directly from a minor or attempting to meet and cause harm to a child in person. These CyberTips also include cases where a child is in apparent imminent danger. https://transparency.meta.com/en-gb/integrity-reports-q1-2024/

children. Over 5.1 million reports related to shared or re-shared photos and videos that contained child sexual abuse material.

We have also partnered with NCMEC to build Take it Down,[44] a global platform for teens who are worried that intimate images they have created might be shared on public online platforms without their consent. Minors can privately generate a hash of their images or videos directly on their own devices, without having to upload their content to the platform. NCMEC then securely houses the hashes, at which point participating tech companies take those hashes and proactively scan for that content on their platforms to remove it.

Whilst the Take It Down platform is for young people, we have also partnered with Revenge Porn Hotline in the UK and international women's safety organisations to develop StopNCII.org,[45] a tool that helps prevent intimate images from being shared on Facebook, Messenger and Instagram. Adults who are concerned their intimate images or videos may be posted on our services can share this content with us in a private, safe and secure channel, and we can then put protections in place to prevent this content from being posted or shared on our apps. We have more than 20 partners on StopNCII.org in the Asia-Pacific region, including the Office of the eSafety Commissioner and experts from Monash University and RMIT.

We also send safety alerts that inform people who have shared CSAM about the harm it can cause, warning that it is against our policies and there are legal consequences for sharing this material. We share these safety alerts in addition to removing the content, banking it and reporting it to NCMEC. We are using insights from these safety alerts to help us identify behavioural signals of those who might be at risk of sharing this material, so we can also educate them on why it is harmful and encourage them not to share it on any surface - public or private.

Youth safety

Creating an experience on Facebook and Instagram that is safe and private for young people, but also fun, comes with competing challenges. In order to make sure we are striking the right balance, we engage closely with experts in this space - and with young people themselves. We have also engaged with parent groups to better understand the resources they need.

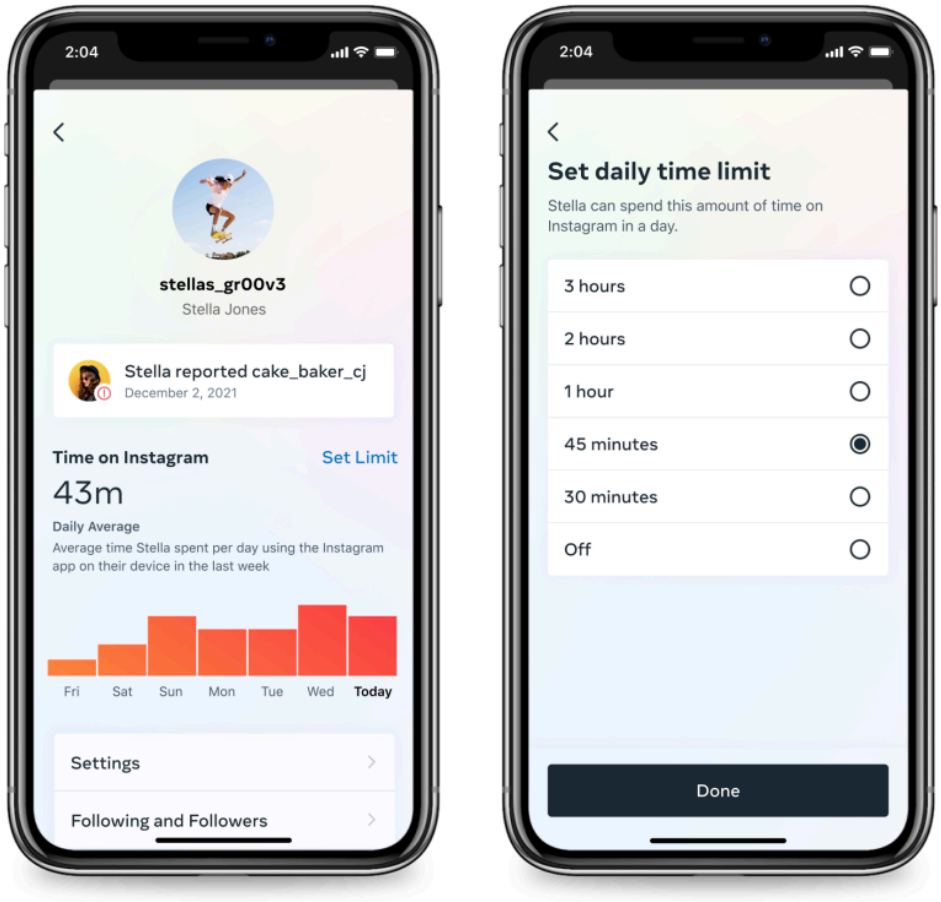
We want people, especially young people, to foster their online relationships in an environment where they feel safe, and where they leave our apps feeling good about the time they spend on them. Our policies prohibit harmful content, or content or behaviour that exploits young people. We work closely with experts in mental health, child psychology, digital literacy and more, to build features and tools so teens can connect online safely and responsibly.

[44] Meta, 'Helping teens avoid sextortion scams' *Newsroom*, 6 February 2024, https://about.fb.com/news/2024/02/helping-teens-avoid-sextortion-scams
[45] StopNCII.org - Stop Non-Consensual Intimate Image Abuse, https://stopncii.org

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 want to provide tools and resources for parents and guardians so they can guide and support their teens.

Since December 2021, we have launched new tools on Instagram and Facebook that provide greater controls for parents. Parents and guardians have been able to view how much time their teens spend on our platforms and set time limits, shown in Figure 4 below.[46] Teens are also able to notify their parents if they report someone, giving their parents the opportunity to talk about it with them. Parents can also approve or deny changes their teens make to their settings, for example to their account privacy settings.

Figure 4: Parent and guardian controls over 'time spent' and reporting



We have also developed a number of resources specifically to provide parents with the details about the tools and features available on our services that assist them in ensuring young people

[46] Meta, 'Raising the standard for protecting teens and supporting parents online', *Newsroom*, 7 December 2021, https://about.fb.com/news/2021/12/new-teen-safety-tools-on-instagram

are having a safe experience, as well as tips and strategies about broader online safety. Two examples of this are:

- **Family Centre.** We have developed resources, accessible from within the apps' supervisory experiences, that include product tutorials and tips from experts, to help parents and guardians discuss social media use with their teens.[47]

- **Parents Portal.** The Parents Portal provides a hub for information and tips on how to help your child navigate their online experience, it also connects parents to online safety organisations around the world that offer additional resources.[48]

- **Parents' Guide to Instagram.** In Australia, we worked with ReachOut to develop a Parents' Guide to Instagram to support parents in better understanding Instagram's safety tools. The Guide contains tips for parents on using Instagram's safety features and on how to have effective conversations with their teens about social media. The Parents' Guide can be downloaded for free on ReachOut's website and we supported ReachOut to publish the Guide and promote it on their social platforms.[49] The Guide was first released in September 2019 and updated in June 2021.[50] We are working with ReachOut to provide a new series of resources for parents in 2024.

Ensuring age-appropriate experiences online

As per our terms, 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.

Meta takes a multi-layered approach to understanding someone's age - we want to keep people who are too young off of Facebook and Instagram, and make sure that those who are old enough receive the appropriate experience for their age.

Understanding a user's age

It is a complex and industry-wide challenge to understand the age of users on the internet. Verifying someone's age is not as easy as it sounds, and relying on identification documentation can raise privacy concerns and may not be truly effective to achieve the intended policy goal.

For this reason, we take a multi-layered approach to understanding a user's age on Facebook or Instagram.

---

[47] Meta, 'Supporting safer and more positive experiences for your family, *Family Centre*, https://familycenter.meta.com/au
[48] Meta, 'Parents', *Safety Center*, https://www.facebook.com/safety/parents
[49] ReachOut, *A parents guide to Instagram*, https://parents.au.reachout.com/-/media/parents/files/pdfs/parents_guide_to_instagram_austrlian_edition2021_reachout.pdf
[50] Meta Policy AU, 'A Parent's Guide to Instagram', Meta Australia Policy Blog, *Medium*, 22 June 2021 (updated 27 January 2023), https://medium.com/meta-australia-policy-blog/a-parents-guide-to-instagram-in-partnership-with-reach-out-30a865e28fcb

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 (ie. 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.

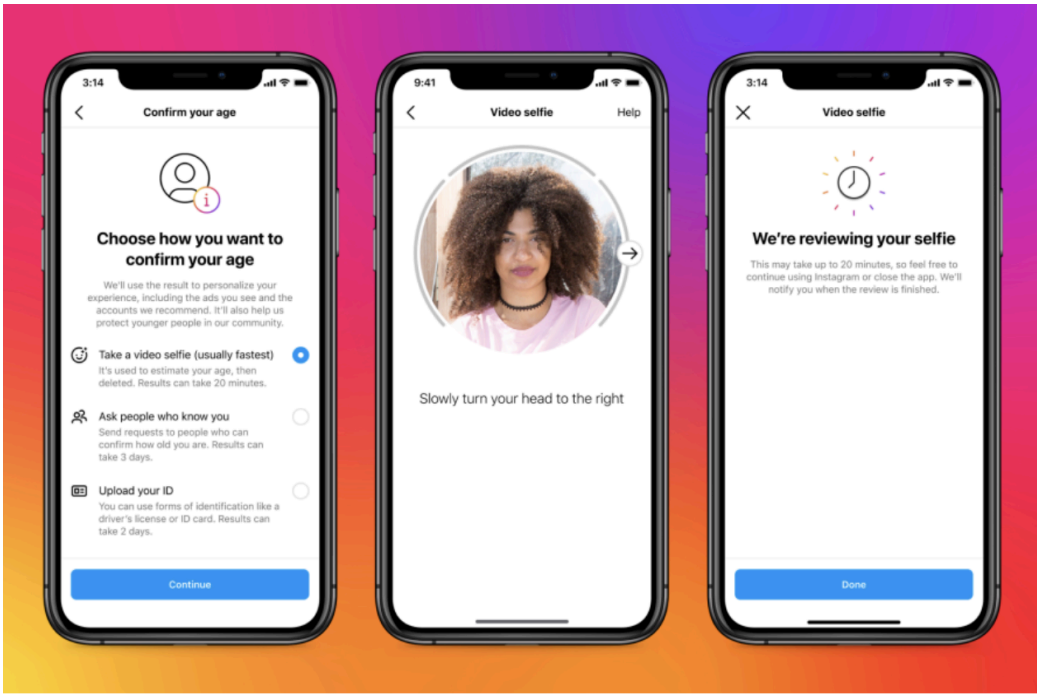
But we also recognise that some people may misrepresent their age online. For that reason, we have been investing in artificial intelligence tools to help us understand someone's real age. Our technology allows us to estimate people's ages, like if someone is below or above 18, using signals. We've focused on using existing data to inform our artificial intelligence technology. Technology like this is new, evolving and it isn't perfect. It also may not always be the most appropriate measure for all use cases. Inaccurate AI predictions could undermine people's ability to use services, for example, by incorrectly blocking them from an app or feature based on false information. Where we do feel we need more information, we have developed a menu of options for someone to prove their age on Instagram and Facebook.

Our team of over 15,000 content reviewers is trained to flag reported accounts that appear to be used by people who are underage. If these people are unable to prove they meet our minimum age requirements, we delete their accounts.

We invest in age prediction models to detect likely teens and ensure they receive age-appropriate experiences for a variety of use cases e.g. restricting adults from sending messages to teens who do not follow them. We use Artificial Intelligence and Machine Learning to help us understand who the youngest members of our community are (specifically 13-18 year olds) - and apply new age-appropriate features we're developing.

We are currently testing two additional options for age-verification,[51] which we have launched in Europe, Mexico, Canada, South Korea, Australia and Japan: (A) Submission of ID or (B) Face-based-age-prediction, offered through Yoti,[52] a 3rd party vendor based out of the UK which provides age estimation services. This is a new option where users upload a video selfie of themselves to verify their age. Face-based age prediction (FBAP) refers to computer-vision systems, which predicts a user's age based on an image of their face. We partner with Yoti because of their industry leading accuracy metrics, their work to minimise bias across skin tones and gender, and their strong privacy guardrails.

[51] Meta, 'Introducing new ways to verify age on Instagram', *Newsroom*, 23 June 2022, https://about.fb.com/news/2022/06/new-ways-to-verify-age-on-instagram
[52] Meta, 'Introducing new ways to verify age on Instagram', *Newsroom*, 23 June 2022, https://about.fb.com/news/2022/06/new-ways-to-verify-age-on-instagram



We are also in discussions with the wider technology industry on how best to share information in privacy-preserving ways that helps apps establish whether people are over a specific age.[53] Globally, we believe requiring app stores to get parents' approval whenever their teens under 16 download apps help us place teens in age-appropriate experiences. By verifying a teen's age in the app store, individual apps would not be required to collect potentially sensitive identifying information. Apps would only need age confirmation from the app store to ensure teens are placed in the right experiences for their age group.

Age-appropriate controls and warnings

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:

- **Defaulting new teen accounts to private.** Wherever we can, we want to stop young people from hearing from adults they don't know, or that they don't want to hear from. We believe private accounts are the best way to do this. In line with this, we now default all new Instagram users who are under the age of 16 in Australia onto a private account. For young people who already have a public account on Instagram, we show them notifications highlighting the benefits of a private account and explaining how to change their privacy settings.

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

[53] A Davis, 'A framework for legislation to support parents and protect teens online', *Medium*, 17 January 2024, https://medium.com/@AntigoneDavis/a-framework-for-legislation-to-support-parents-and-protect-teens-online-6565148b26b1

Location is defaulted off. On Instagram and Messenger, adults are unable to start a conversation with a teen who is not connected with them. Over and above this, we prevent teens from messaging potentially suspicious adults (e.g. who may have been blocked or reported by other teens) they are not connected with, and also prevent teens from receiving messages from all accounts they are not connected with by default.

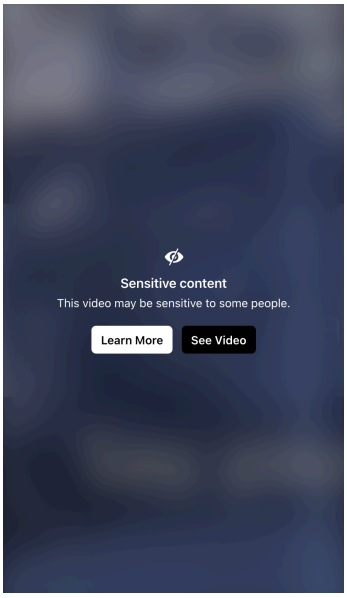
- **Limiting advertisers' ability to reach young people.** We only allow advertisers to target ads to people under 18 based on their age and location. This means that previously available targeting options, like those based on interests or on their activity on other apps and websites, are not available 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).

- **Warning label for sensitive content.** There are 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).
    - 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.

  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 in Figure 7 below. The content will not appear, or present the option of viewing it, for users who are under the age of 18.

Figure 7: Example of a piece of content that is "marked as sensitive" on Facebook



- **Safety Notices in messaging.** In addition to the restrictions explained above, we also send safety notices to users in Messenger and Instagram, if we believe an adult could be pursuing a potentially inappropriate private interaction with a teen.[54]

- **Making it more difficult for adults to find and follow teens.** We have developed new technology that allows us to find accounts that have shown potentially suspicious behaviour and stop those accounts from interacting with young people's accounts. By "potentially suspicious behaviour", we mean accounts belonging to adults that may have recently been blocked or reported by a young person, for example.

  Using this technology, we do not 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 will not 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. The reverse also holds - that is, teen accounts' ability to interact with such accounts are similarly restricted. We will continue to look for additional places where we can apply this technology.

Mental health and wellbeing

Being socially connected, both online and offline, plays an important role in our mental health and wellbeing. We believe our platforms have a responsibility to not only provide a safe

[54] J Sullivan, 'Preventing unwanted contacts and scams in Messenger', *Messenger News*, 21 May 2020, https://messengernews.fb.com/2020/05/21/preventing-unwanted-contacts-and-scams-in-messenger

environment but to also support people in any time of need. We want the services that Meta provides to be a place for meaningful interactions with your friends and family - enhancing people's relationships offline, not detracting from them.

We also recognise that people's time spent online should be balanced, positive and age appropriate, and so we invest heavily in the following areas so that a user's time spent on our services is positive and purposeful:

- **Research.** We have a dedicated team of researchers and support global and local research in Australia to understand the impact of social media, mental health and wellbeing.
- **Partnerships.** As mentioned above, Meta has convened a global Safety Advisory group. We have also developed strong relationships with global and local organisations to ensure our programs and tools are fit for purpose for Australians.
- **Tools and resources.** We have created a number of tools and resources, informed by our research and partnerships, to enable positive experiences, and guide users through finding support. These are outlined in more detail below.

Approach to eating disorder content

We have developed - and continue to review and update - our approach to eating disorder content in consultation with experts around the world. Our specific policies about eating disorder content aim to strike a balance between preventing people from seeing harmful, sensitive or upsetting content and giving people space to talk about their own experiences, which experts say is important. We do not allow content that promotes, encourages or glorifies eating disorders and we remove it as soon as we become aware of it. We also have a dedicated in-platform reporting option for eating disorder content.

While it can be challenging to proactively identify eating disorder content, due to the many different forms this can take, between January and March 2024, we found and took action on more than 99.4% of the suicide, self-harm and eating disorder content on Facebook and Instagram before it was reported to us.[55]

Recent tools that we have introduced include nudging teens towards other topics if they have been scrolling on the same topic on Instagram for a while.[56] When someone searches for, or posts, content related to eating disorders or body image issues, they will see a pop-up with tips and an easy way to connect to organisations offering support, including the Butterfly

[55] Meta, 'Community Standards Enforcement Report - Suicide and Self-Injury', *Transparency Center*, https://transparency.meta.com/reports/community-standards-enforcement/suicide-and-self-injury/facebook/
[56] Instagram, 'New tools and resources for parents and teens in VR and on Instagram', 14 June 2022, https://about.instagram.com/blog/announcements/tools-and-resources-for-parents-and-teens-in-vr-and-on-instagram

Foundation in Australia.[57] We have also updated this message to make it much more prominent, removing friction by reducing the number of click-throughs to get to helplines and ability to call the helpline within these resources pop-ups.

Approach to suicide and self-injury

We regularly consult with experts in suicide and self-injury to help inform our policies and enforcement, and work with organisations around the world to provide assistance to people in distress.

We define self-injury as the intentional and direct injuring of the body, including self-mutilation and eating disorders. We remove any content that encourages suicide or self-injury, including fictional content such as memes or illustrations and any self-injury content which is graphic, regardless of context. We also remove content that identifies and negatively targets victims or survivors of suicide or self-injury seriously, humorously or rhetorically, as well as real time depictions of suicide or self-injury.

We allow people to discuss topics relating to suicide and seslf-injury because we want Facebook and Instagram to be spaces where people can share their experiences, raise awareness about these issues and seek support from one another. However, we make content about recovery from suicide of self-harm that is allowed on our services harder for teens to find.[58]

On both Facebook and Instagram, we use machine learning and image-based technology to proactively identify and take action on potential suicide and self-injury content (either by removing it automatically or escalating it to human reviewers to take appropriate action) and expand our ability to get timely help to people in need.

We also work with experts in suicide prevention and safety to develop support options for people posting about suicide. Experts say that one of the best ways to help prevent a suicide is for people in distress to hear from others who care about them. Meta has a role to play in connecting people in distress with people who can offer support.

We have released suicide prevention support on Facebook Live and introduced artificial intelligence to detect posts that indicate someone may be at risk of imminent harm. And when there's risk of imminent harm, we work with emergency responders who can help. We also connect people more broadly with mental health resources, including support groups on Facebook.[59]

---

[57] Instagram, 'How we're supporting people affected byeating disorders and negative body image', 23 February, 2021, https://about.instagram.com/blog/announcements/how-were-supporting-people-affected-by-eating-disorders-and-negative-body-image

[58] 'New protections to give teens more age-appropriate experiences on our apps', https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps/

[59] Meta, 'Getting our community help in real time', *Newsroom*, 27 November 2017, https://about.fb.com/news/2017/11/getting-our-community-help-in-real-time

Research

We have a dedicated team of researchers that work to understand the impact of social media on mental health. We employ social psychologists, social scientists and sociologists, and we collaborate with top scholars to better understand wellbeing and the impact of social media on mental health.

According to the research, the impact of technology on senses of wellbeing depend on how people use it.

In general, when people spend a lot of time passively consuming information — reading but not interacting with people — they report feeling worse afterward. However, actively interacting with people — especially sharing messages, posts and comments with close friends and reminiscing about past interactions — is linked to improvements in wellbeing.[60]

Moira Burke, Meta's Data Scientist and Wellbeing Researcher, has undertaken a number of studies on the intersection of wellbeing and social technology.[61] These studies found that people tend to have higher quality interactions on social media with their strong personal ties, such as friends, family and romantic partners. Further, a study we conducted with Robert Kraut at Carnegie Mellon University found that people who sent or received more messages, comments and Timeline posts reported improvements in social support, depression and loneliness. The positive effects were even stronger when people talked with their close friends online.[62]

We've used this research to inform user experiences online by introducing changes to News Feed, and tools such as the Activity Dashboard, suicide prevention tools, hiding likes, and the 'Take a Break' tool (all discussed below).

We made these important changes because we want to support wellbeing through meaningful interactions, even if it decreases time spent on the platform. In fact, shortly after we made the Meaningful Social Interactions change to News Feed in 2018, we saw time spent on the platform go down by 50 million hours per day.

Additional well-being tools

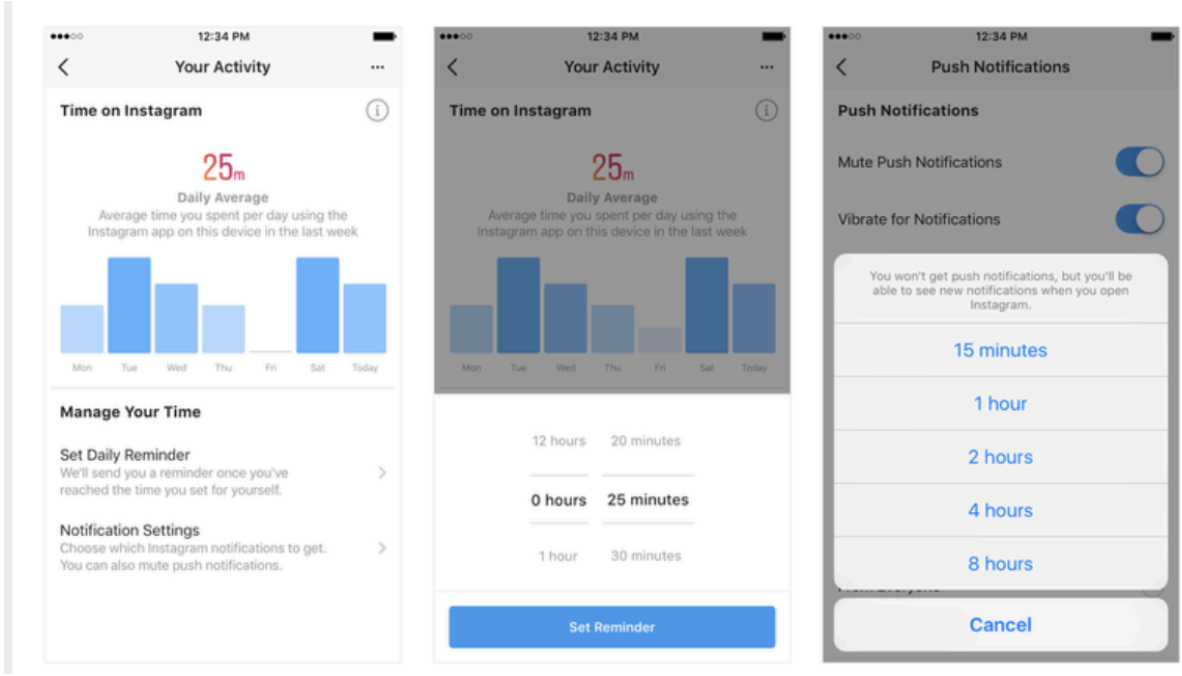
We want the time people spend on Facebook and Instagram to be intentional, positive and inspiring, and we have developed tools to help users understand how much time they spend on our platforms so they can better manage their experience. These include:

---

[60] P Verduyn, et al., 'Do social media sites enhance or undermine subjective wellbeing? A critical review', *Social Issues and Policy Review*, 13 January 2017, https://spssi.onlinelibrary.wiley.com/doi/full/10.1111/sipr.12033
[61] M Burke, *Research,* https://research.facebook.com/?s=burke+moira
[62] Meta, 'Hard questions: Is spending time on social media bad for us?' *Newsroom*, 15 December 2017, https://about.fb.com/news/2017/12/hard-questions-is-spending-time-on-social-media-bad-for-us

- **Improving Feed quality.** As mentioned above, we've made several changes to Feed to provide more opportunities for meaningful interactions, and reduce passive consumption of low-quality content.[63] We demote things like clickbait headlines and false news. We optimise ranking so posts from the friends you care about most are more likely to appear at the top of your feed. Similarly, our ranking promotes posts that are personally informative. We also redesigned the comments feature to foster better conversations.
- **Activity Dashboard.** The Activity Dashboard, shown in Figure 9 below, was introduced in 2018 to help people manage their time on Facebook and Instagram. The Dashboard allows people to see the average time spent on the app , and allows them to set reminders once they've reached the amount of time they want to spend on the app.[64]

Figure 9: Activity Dashboard



- **Hide Likes on Facebook and Instagram.** We tested hiding like counts to see if it might depressurise people's experience on Instagram.[65] What we heard from people and experts was that not seeing like counts was beneficial for some and annoying to others, particularly because people use like counts to get a sense of what's trending or popular. We now give users the option to hide like counts on all posts they see in their feed. They also have the option to hide like counts on their own posts, so others can't see how many likes their posts get.

[63] M Zuckerberg, Meaningful social interaction post, *Facebook,* 2 November 2017, https://www.facebook.com/zuck/posts/10104146268321841
[64] Meta, 'New tools to manage your time on Facebook and Instagram', *Newsroom*, 1 August 2018, https://about.fb.com/news/2018/08/manage-your-time
[65] Meta, 'Giving people more control on Instagram and Facebook', *Newsroom*, 26 May 2021, https://about.fb.com/news/2021/05/giving-people-more-control

- **Take a Break.** In December 2021, we announced a new tool called Take a Break which will empower people to make informed decisions about how they're spending their time.[66] If someone has been scrolling for a certain amount of time, we'll ask them to take a break from Instagram and suggest that they set reminders to take more breaks in the future. We'll also show them expert-backed tips to help them reflect and reset.

  We're encouraged to see that teens are using Take A Break. Early test results show that once teens set the reminders, more than 90 per cent of them keep them on. This tool has been commended by experts and researchers. Boris Radanoic from UK Safer Internet Centre said "we welcome Instagram's new Take A Break feature, which we hope will be a meaningful way to encourage healthy social media use, particularly among younger users. Whilst taking regular breaks from screens has been challenging recently, it has been good advice for many years, and initiatives that encourage this are to be supported. We will continue to work with Instagram in this regard and hope that this represents a step in the right direction."

We offer a number of online Centres that work as a centralised source of authoritative, up to date information for users. This includes a Safety Centre that provides resources on online wellbeing.[67]

Combatting sextortion

Recognising the growing trend of sextortion, including financially-motivated sextortion, we recently announced an updated suite of measures designed specifically to better protect young people from this harm type. Specifically, we recently expanded on our existing safety mitigations:[68]

- We added links to new child safety helplines in our education pop-up to users (discussed above) about how to block and report (including for nudity/sexual activity, sexual exploitation, and sharing private images).
- We are now showing pop-up warnings to teens that have recently engaged with an account we've disabled for sextortion (these warnings will have sextortion-specific messaging and direct victims to relevant resources).
- We have developed signals to detect potential sextorters and prevent them from finding or interacting with teen accounts (e.g., removing the "message" button from their IG profile, routing their message requests to Hidden Folders, stopping them seeing teens in

[66] Meta, 'Raising the standard for protecting teens and supporting parents online', *Newsroom*, 7 December 2021, https://about.fb.com/news/2021/12/new-teen-safety-tools-on-instagram
[67] *See* https://facebook.com/safety
[68] Meta, 'New Tools to Help Protect Against Sextortion and Intimate Image Abuse', *Newsroom*, 11 April 2024, https://about.fb.com/news/2024/04/new-tools-to-help-protect-against-sextortion-and-intimate-image-abuse/

following/follower and like lists, and preventing them from being able to find teen accounts in Search).

Another new feature that we are testing leverages technology on device, which, when turned on, will automatically analyse when images containing nudity are being sent and/or received in Instagram DMs. If the feature is turned on, when a receiver receives a nude image, the app will cover the image with a warning screen and show a pop-up message reminding the receiver of our safety features, and that they shouldn't feel pressured to respond; when a sender has sent a nude image, the app will cover it with a warning screen reminding the sender to be careful, that people can screenshot and re-share these images, and that they can unsend the image if they've changed their mind.

The feature will be default-on for teens, and because the feature leverages machine learning to analyze these images on the device itself, nudity protection will work in fully end-to-end encrypted chats – where neither Meta nor any other third party can access these images (unless someone chooses to report them to us).

Combatting hate speech

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.[69]

We also know that there is not always a clear consensus in defining a piece of content as hate speech because different content may have different meanings, or the intent or context behind them may be unclear. This is why Meta also provides an appeals process for people to tell us when they think we have made the wrong decision on content that we enforce on.

We continue to receive feedback from partners globally as well as in Australia on emerging risks and move quickly to address them. In 2020, we established an Australia-specific Combatting Online Hate Speech Advisory Council, which shares important advice and insights that enable us to shape and strengthen our policy and programmatic responses to hate speech content on

[69] Meta, 'Facebook Community Standards - Hate Speech', *Transparency Center*, https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

our services, specifically targeting Australians. The group comprises representatives of minority or marginalised communities and academic experts in different forms of online hate.

We also partner with Australian organisations and experts to support community initiatives and research relating to combatting hate speech. Some of our most recent efforts include:

- **Online Hate Prevention Institute**: In 2023, we supported Australia's Online Hate Prevention Institute to conduct analysis into hate speech targeting First Nations Australians. This built on our support to OHPI to conduct research investigating Anti-Asian Hate on social media that was published in late 2022. We have also been working with OHPI to ensure that our policies and enforcement on anti-Semitic content are working effectively.
- **Islamophobia Register Australia (IRA)**: We supported the 4th Islamophobia report (2022) to shed light on the ongoing threat faced by Muslim individuals in their daily lives, conducted by the register in partnership with Charles Sturt University and the Islamic Sciences and Research Academy. The analysis of these reported cases serves as a valuable source for understanding the manifestations of Islamophobia within the Australian context. In 2023, we also provided funding to IRA to deliver a series of workshops in 2024 on navigating Islamophobia to Muslim youth aged 15-30 years across Australia.
- **Media Diversity Australia:** In May 2023, we supported Media Diversity Australia to launch their research[70] investigating the online abuse of diverse journalists and media workers in Australia. The Australia-first research aimed to identify, understand and address online abuse and harassment experienced by diverse journalists and media workers, including those from Indigenous, CALD and LGBTQI+ communities.

Combatting violent and extremist content

Our policies prohibit the sharing of violent extremist and terrorism content and we use a mixture of automation and human review to enforce these policies. In the first quarter of 2024, we removed 8.4 million pieces of content, 99.3% of which we actioned proactively.[71] We also leverage the benefit of industry collaboration through the Global Internet Forum to Counter Terrorism (GIFCT). To disrupt and prevent the spreading of identified terrorist and violent extremist content online, GIFCT and its members continued to strengthen the GIFCT's hashsharing database. The number of hashes grew to approximately 390,000 distinct items of terrorist and violent extremist content in the forms of images, videos, and texts.[72]

[70] Media Diversity Australia, 'Research & Resources - Online Safety of Diverse Journalists', https://www.mediadiversityaustralia.org/online-safety-of-diverse-journalists/
[71] Meta, 'Community Standards Enforcement Report - Dangerous Organizations: Terrorism and Organized Hate', *Transparency Center*, https://transparency.meta.com/reports/community-standards-enforcement/dangerous-organizations/facebook/
[72] GIFCT, *2023 GIFCT Annual and Transparency Report*, https://gifct.org/wp-content/uploads/2024/04/GIFCT-Annual-Report-2023.pdf

In addition to our policies and technology, a key approach for Meta in dealing with violent extremist content is through various product interventions. The DOI Search Intercept is an intervention that is activated when users search for DOI-related terms on platforms such as Facebook, Instagram, and Threads. It aims to prevent access to harmful content, redirect at-risk users to support services, and provide educational resources. This measure complements other efforts to reduce the visibility of content that breaches DOI policies.

In Australia, we have also rolled out the Search Redirect program. It responds to region-specific terms associated with DOI concerns, such as terrorism and extremism, in collaboration with a local delivery partner. In the case of Australia, we have partnered with the New South Wales Department of Communities and Justice where individuals who choose to seek assistance from the delivery partner are then redirected to the landing page of the partner where an array of help services are provided to these individuals who are seeking to leave violent extremism. Beyond the country specific program, specialized responses are also developed to deal with challenges posed by QAnon and tackle Holocaust Denial.

Combatting disinformation & misinformation

Meta is committed to create a place for expression and give people a voice, and to the integrity of our platforms. We work to combat coordinated inauthentic behaviour that can lead to misinformation and disinformation on our services. We do so via a combination of: (1) policies and enforcement to remove and reduce mis- and disinformation and inauthentic activity on our services; (2) products and programs to give people greater context about what they are seeing on our services and to promote greater media literacy; (3) transparency tools and research.

To provide transparency about this investment, Meta is a founding signatory of the DIGI Australian Disinformation and Misinformation Industry Code.

Coordinated inauthentic behaviour

We have policies that prohibit inauthentic activity on our services and we enforce these using a mixture of automation and human review. In the social media landscape and beyond, foreign interference relies on inauthenticity - where users misrepresent themselves, through fake profiles or non-transparent behaviours - and coordination.

We consider authentic communications as a central part of people's experience on Meta's services. People find value in connecting with their friends and family, and they also find value in receiving updates from the Pages and organisations that they choose to follow. For this reason, authenticity has long been a requirement of our Community Standards.

Our policies in this space have been through a number of iterations over recent years, to reflect our deepening understanding of the phenomenon of inauthentic behaviour.[73] We have an Inauthentic Behaviour policy, which has a number of components:

- *Coordinated Inauthentic Behaviour* (CIB). We define this as groups of accounts and Pages that work together to mislead people about who they are and what they are doing. We use a combination of policies, tools, expert teams and partnerships to detect and remove networks of inauthentic behaviour (IB) and CIB - both foreign and domestic.

- *Foreign or Government Interference.* These are either (1) foreign-led efforts to manipulate public debate in another country in a way that is inauthentic; or (2) inauthentic behaviour operations run by a government to target its own citizens. If we see any of these instances, we will apply the broadest enforcement measures, including the removal of every on-platform property connected to the operation itself and the people and organisations behind it.

- *Other inauthentic behaviour*, including financially-motivated activity like spam or fake engagement tactics that rely on inauthentic amplification or evading use of enforcement (separate to use of fake accounts). The full list of tactics that we do not allow is available as part of our Community Standards.[74] We enforce against other inauthentic behaviour based on specific protocols that may involve temporary restrictions, warnings, down-ranking in Facebook News Feed, or removal.

We have invested significantly in a team that is able to detect the various forms of Inauthentic Behaviour on our services.

We regularly report on our efforts to disrupt CIB through our Community Standards Enforcement Report and Quarterly Adversarial Threats report. As mentioned in our most recent report, Meta has removed over 200 covert influence operations between 2017-2022.[75] In 2023, more than half of these CIB networks targeted audiences outside of their countries of operation.[76]

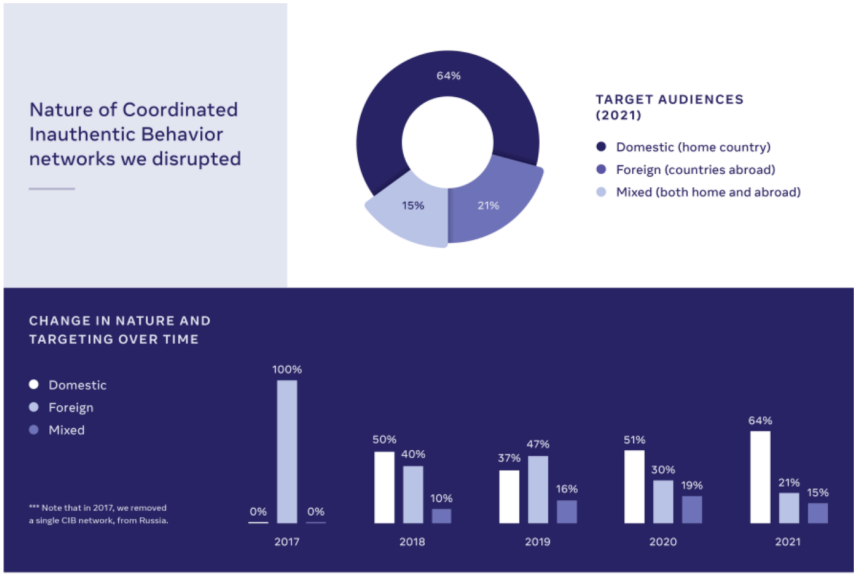
Since 2017, Meta has taken action on five instances of CIB operations that targeted Australians. We removed the majority of these networks before they were able to build authentic audiences. In our 2024 transparency report under the misinformation and disinformation code, we reported on trends in CIB in 2023, including an increase in China based CIB disruptions and the continued

[73] Meta, 'How we respond to Inauthentic Behaviour - policy update', *Newsroom*, 21 October 2019, https://about.fb.com/news/2019/10/inauthentic-behavior-policy-update
[74] Meta, 'Facebook Community Standards - Inauthentic Behavior', *Transparency Center*, https://www.facebook.com/communitystandards/inauthentic_behavior
[75] Meta, 'Recapping our 2022 coordinated inauthentic behaviour enforcements', *Newsroom*, 15 December 2022, https://about.fb.com/news/2022/12/metas-2022-coordinated-inauthentic-behavior-enforcements
[76] Meta, 'Recapping our 2022 coordinated inauthentic behaviour enforcements', *Newsroom*, 15 December 2022, https://about.fb.com/news/2022/12/metas-2022-coordinated-inauthentic-behavior-enforcements

role of for-hire surveillance organisations in covert influence operations globally, with many of the operations covered in our 2023 Quarterly Adversarial Threats reports being attributed to private entities.[77]

This is a highly adversarial space where deceptive campaigns we take down continue to try to come back and evade detection by us and other platforms, which is why we continuously take action as we find further violating activity.

Target of Coordinated Inauthentic Behaviour Disruptions, 2017 - 2022[78]



We know that CIB threats are rarely confined to one platform. We share our findings and threat indicators with industry peers so they too can detect and stop threat activity, and we can build our collective response to CIB.

Fake accounts

We do not allow fake accounts on Facebook and Instagram, as they can be vehicles for a range of harmful content and behaviour, including spreading misinformation. In the first quarter of 2024, we took action on 631 million fake accounts on Facebook, 99.4% of which we detected proactively.[79] The majority are caught within minutes of registration.

---

[77] Meta, 'Meta response to the Australian Code of Practice on Disinformation and Misinformation', May 2024, https://digi.org.au/wp-content/uploads/2024/05/Meta-Transparency-Report-2024-Australian-Code-of-Practice-on-Disinformation-and-Misinformation.pdf

[78] We define targets as:

- Domestic: IO that targets public debate in the same country from which it operates.
- Foreign: IO that targets the public debate in a different country from which it operates.
- Mixed: We also see IO campaigns and threat actors that run campaigns that target both domestic and foreign audiences

[79] Meta, 'Community Standards Enforcement Report - Fake Accounts', Q1 2024, https://transparency.fb.com/data/community-standards-enforcement/fake-accounts/facebook

Misinformation

Meta is committed to stopping the spread of misinformation. We use a combination of enforcement technology, human review and independent fact checkers to identify, review and take action on this type of content. Our strategy for misinformation comprises three main pillars: remove, reduce and inform.

Remove

Generally, our Community Standards[80] and Ad Standards[81] apply to all content, including content generated by AI, and we will take action against this type of content when it violates these policies.

- **Misinformation and harm**. We remove misinformation where it is likely to directly contribute to the risk of imminent physical harm.

  In determining what content constitutes misinformation in this category, 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 with a presence on the ground in a country to determine the truth of a rumour about civil conflict, and partnering with health organisations during the global COVID-19 pandemic.

- **Election-related misinformation that may constitute voter fraud and/or interference.** Under our policies, we remove content that is likely to directly contribute to interference with the functioning of political processes. This includes misinformation about the dates, locations, times, and methods for voting or voter registration (for example: claims that you can vote using an online app), and misinformation about who can vote, qualifications for voting, whether a vote will be counted, and what information or materials must be provided in order to vote.[82]

  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.

- **Violence-Inducing Conspiracy Theory policy.** Our dangerous organisations policy captures content relating to "violence-inducing conspiracy theories". As of September 2021, we identified and removed over 1,013 militarised social movements on our platforms and in total, removed about 7,900 Pages, 31,900 groups, 830 events, 105,000

[80] Meta, 'Facebook Community Standards', *Transparency Center,* https://transparency.fb.com/policies/community-standards
[81] Meta, 'Introduction to the Advertising Standards', https://transparency.fb.com/policies/ad-standards
[82] Meta, 'Facebook Community Standards - Misinformation', *Transparency Center,* https://transparency.meta.com/policies/community-standards/misinformation

Facebook profiles and 40,800 Instagram accounts. Some of these Pages, groups, events, profiles and accounts were located in Australia.[83]

Reduce and inform

For content that does not violate our Community Standards but is rated as false or altered by Meta's independent third-party fact-checking partners, we significantly reduce the number of people who see it through a number of measures. We believe that public debate and democracy are best served by allowing people to debate different ideas, even if they are controversial or wrong - but we take steps to limit the distribution of misinformation that has been found to be false by independent, expert fact checkers.

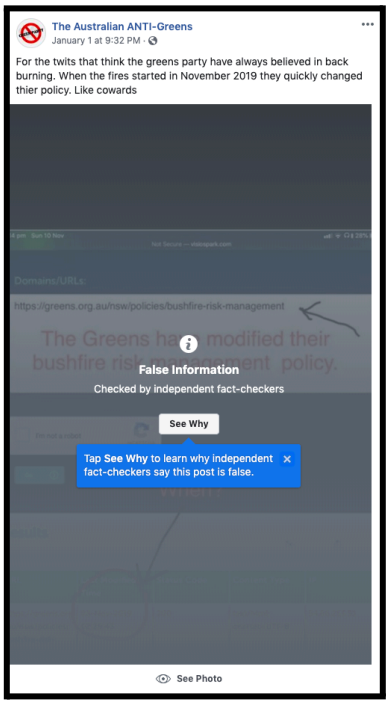
- **Third-party fact-checking program.** Meta partners with third-party fact-checking organisations globally, to assess the accuracy of content on our services. We have commercial arrangements with independent third-party fact-checking organisations for them to review and rate the accuracy of posts on Facebook and Instagram.

  Since 2016, we have built the largest global fact-checking network of any platform and have contributed more than $150 million to programs supporting our fact-checking efforts.[84] We now have over 90 partners around the world to review and rate viral misinformation in more than 60 languages.

  In Australia, we partner with Australian Associated Press, Agence France Presse and RMIT FactLab.[85] All fact-checks by these partners are publicly available on their websites.[86] In preparation for the 2023 Aboriginal and Torres Strait Islander Voice Referendum, we provided a one-off funding boost to AAP and AFP, so that they could increase their capacity in the lead up to the referendum.[87]

- **Warning labels.** Once a third-party fact-checking partner rates a post as 'false', we apply a warning label that indicates it is false and shows a debunking article from the fact checker. It is not possible to see the content without clicking past the warning label.

[83] Meta, 'An update to how we address movements and organizations tied to violence', *Newsroom,* 19 August 2020 (updated 9 November 2021), https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence
[84] Meta, 'An Update on Facebook News', *Newsroom*, 29 February 2024, https://about.fb.com/news/2024/02/update-on-facebook-news-us-australia
[85] Meta, 'A list of our independent fact-checking partners, by country', https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking/partner-map
[86] Agence France Presse Australia, *Fact Check*, https://factcheck.afp.com/afp-australia; Australian Associated Press, *AAP Fact Check*, https://www.aap.com.au/category/factcheck/; RMIT, *RMIT FactLab*, https://www.rmit.edu.au/about/schools-colleges/media-and-communication/industry/factlab/debunking-misinformation
[87] Meta, 'How Meta is preparing for the Voice to Parliament Referendum', *Medium*, 9 July 2023 https://medium.com/meta-australia-policy-blog/how-meta-is-preparing-for-the-voice-to-parliament-referendum-282632baccfb



We also notify people before they try to share this content or if they shared it in the past.

AI-generated content is also eligible to be reviewed and rated by our independent fact-checking partners. One of the rating options is Altered, which includes, "Faked, manipulated or transformed audio, video, or photos."[88]

- **Ensuring that fewer people see false information.** Once found to be false or altered, Meta reduces the distribution of that content so it appears lower in Feed, which slows its distribution significantly. And on Instagram, we remove it from recommendable surfaces like Explore and hashtag pages, and downrank content in Feed and Stories. We also do not allow an ad to run if it has been debunked.

- **Searching for content that makes claims debunked by our fact-checking partners, to apply the same treatments**. Based on one factcheck, our technology is able to identify duplicates of debunked stories and limit the distribution of similar posts. In April 2020 alone, we applied the label and reduced the distribution of more than 50 million posts worldwide, based on more than 7,500 fact-checks.[89]

- **Taking action on Pages, Groups, accounts, or websites found to repeatedly share misinformation, including removing them from recommendations.** When Pages, Groups

[88] Meta, 'How Meta Is Preparing for the EU's 2024 Parliament Elections', *Newsroom*, 25 February 2024, https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections
[89] Meta, An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19, *Meta Newsroom,* 16 April 2020 (updated 26 May 2021), https://about.fb.com/news/2020/04/covid-19-misinfo-update

or websites repeatedly share content that's been debunked by fact-checking partners, they will see their overall distribution reduced, and will lose the ability to advertise or monetise within a given time period. We will also let people know if they are about to join a group that has Community Standards violations, so they can make a more informed decision before joining.[90] If they continue to share misinformation, the Page or Group is removed in its entirety. This includes Pages operated by public figures.

We know our third-party fact-checking program is working and people find value in the warning screens we apply to content after a fact-checking partner has rated it. We surveyed people who had seen these warning screens on-platform and found that 74% of people thought they saw the right amount or were open to seeing more false information labels, with 63% of people thinking they were applied fairly.[91]

Australian Disinformation and Misinformation Industry Code

As outlined above, Meta is a founding signatory of the DIGI Australian Disinformation and Misinformation Industry Code.

Our 2024 transparency report under the code[92] outlines the steps we took during the 2023 calendar year to meet the 38 commitments that we opted into over that reporting period and includes:

- **Global adversarial threats**. In 2023, we disrupted a number of Co-ordinated Inauthentic Behaviour (CIB) networks. More than half of these CIB networks targeted audiences outside of their countries of operation. We removed the majority of these networks before they were able to build authentic audiences. We identified four trends in CIB throughout 2023, specifically: an increase in China-based CIB disruptions, for-hire surveillance operations being behind CIB globally, abuse of domain name infrastructure and the Russian network 'Doppelgänger' trying to stay online.

- **Australia-specific data on our misinformation efforts**. Between 1 January and 31 December 2023, we:
    - Took action on over 9,700 pieces of content across Facebook and Instagram in Australia for violating our Misinformation policies. In addition to this, over 6,200 ads were removed in Australia for violating our Misinformation policy.
    - Displayed warnings on over 9.2 million distinct pieces of content on Facebook, and over 510,000 on Instagram, in Australia (including reshares) based on articles written by our third-party fact-checking partners.

[90] Meta, 'Changes to keep Facebook Groups safe', *Newsroom*, 17 March 2021 (updated 20 October 2021), https://about.fb.com/news/2021/03/changes-to-keep-facebook-groups-safe
[91] Meta, 'How Meta's third-party fact-checking program works', 1 June 2021, https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works
[92] Meta, 'Meta response to the Australian Code of Practice on Disinformation and Misinformation', May 2024, https://digi.org.au/wp-content/uploads/2024/05/Meta-Transparency-Report-2024-Australian-Code-of-Practice-on-Disinformation-and-Misinformation.pdf

    - Removed over 75,000 ads in Australia for not complying with our Social Issues, Elections and Politics (SIEP) ads policy.

- **Invested in media literacy research and events**, including:
    - Sponsoring and hosting the Australian Media Literacy Alliance's Australian Media Literacy Summit, which brought together a range of journalists, academics, educators, librarians and other experts to discuss and learn about diverse ways to strengthen media literacy education in Australia.
    - Supporting Western Sydney University's 2023 Young People and News longitudinal survey, which provides findings about the news attitudes, practices and experiences of young Australians aged between 8-16 years.

Our 2024 transparency report also included a case study outlining our comprehensive strategy leading up to Australia's 2023 Aboriginal and Torres Strait Islander Voice Referendum to proactively detect and remove content that breached our services, combat misinformation and harmful content, and promote civil participation. Our efforts included:

- the development of referendum day reminders on Facebook and Instagram to encourage people to vote
- increasing the capacity of our third-party fact-checkers Australian Associated Press (AAP) and Agence France Presse through a one-off funding boost ahead of the Referendum
- partnering with AAP to empower people to identify false news and disinformation through a new media literacy awareness campaign
- supporting academic analysis relating to information integrity surrounding the referendum debate and its outcomes, including in partnership with RMIT Cross-Check and La Trobe University.

Programs and partnerships

We support and deliver a range of programs and partnerships that aim to educate and inform Australian users on how to identify and avoid misinformation and invest in research to contribute to wider understanding about, and future efforts to address, evolving trends relating to misinformation.

Some of our most recent programs and partnerships work in Australia includes:

- In the lead up to the 2023 Australian Aboriginal and Torres Strait Islander Voice Referendum, we launched a new media literacy campaign with the Australia Associated Press building on our "Check The Facts" campaign which ran ahead of the 2022 Federal Election in October 2021 and early 2022. The campaign ran for 6 weeks with a combined

reach on Facebook and Instagram in Australia of over 10 million users, creating over 40 million impressions.[93]

- Sponsoring and hosting the Australian Media Literacy Alliance's Australian Media Literacy Summit in March 2023,[94] which brought together a range of journalists, academics, educators, librarians and other experts to discuss and learn about diverse ways to strengthen media literacy education in Australia.
- Supporting the Western Sydney University's 2023 Young People and News longitudinal survey, which provides findings about the news attitudes, practices and experiences of young Australians aged between 8-16 years.
- In July 2023, we developed a training series called 'Connect, Alert, Inform', which aimed to strengthen emergency communicators' skills in using Meta's services to more effectively deliver critical disaster-related information. It included a module delivered by RMIT CrossCheck on addressing and avoiding the amplification of misinformation surrounding a disaster situation.
- We supported analysis relating to information integrity surrounding the Voice Referendum debate and its outcomes, specifically:
    - RMIT CrossCheck's work to promote accurate and corrective information on the referendum debate and boost media literacy; and
    - La Trobe University's report 'Influencers and Messages: Analysing the 2023 Voice to Parliament Referendum Campaign',[95] which examined the main topics of the debate, key actors and campaign strategies for Yes and No, and the prevalence and influence of misinformation and disinformation.

Transparency and research tools

We also recognise the importance of supporting transparency efforts to encourage scrutiny of online misinformation trends. We support transparency efforts via: our own products; and by supporting research by academics and experts.

We have built industry-leading products like the Meta Ad Library, which is a searchable archive of all social issues and political ads on our services in Australia. We have progressively added functionality and real-time data on these ads.

The Meta Content Library and Content Library API 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 in Australia. Meta has partnered with the Inter-university Consortium for Political and Social Research (ICPSR) at the

[93] Meta, 'Meta response to the Australian Code of Practice on Disinformation and Misinformation', May 2024, https://digi.org.au/wp-content/uploads/2024/05/Meta-Transparency-Report-2024-Australian-Code-of-Practice-on-Disinformation-and-Misinformation.pdf
[94] Australian Media Literacy Alliance, 'Australian Media Literacy Summit 2023', https://medialiteracy.org.au/summit
[95] A Carson, et al, 'INFLUENCERS and MESSAGES: Analysing the 2023 Voice to Parliament Referendum Campaign', 17 April 2024, https://opal.latrobe.edu.au/articles/report/INFLUENCERS_and_MESSAGES_Analysing_the_2023_Voice_to_Parliament_Referendum_Campaign/25604352

University of Michigan to share public data from Meta's platforms in a responsible, privacy-preserving way. Researchers can apply for access to these tools via the ICPSR.[96]

Election integrity

We recognise that it is important to have a specific strategy in place in advance of elections such as the 2022 election and similar events such as the recent "Voice to Parliament" referendum. In advance of each of these, we publish our strategy to combat misinformation, voter interference and potentially harmful content on our platforms.[97] We also ensure that we have strong partnerships with local regulators and law enforcement to deliver our efforts. In 2022, for example, we worked closely with the Australian Electoral Commission (AEC), the Government's Election Integrity Assurance Taskforce (EIAT), and a range of government and law enforcement agencies in the lead up to the election.

This involved working closely with the AEC to respond to all content where they had concerns about compliance with Australian electoral law. We also worked with the Australian Government's EIAT to undertake scenario planning for different online issues that may arise during the course of an election campaign.

In the lead-up to recent elections, we have also run prompts to encourage people to vote. In 2023, we deployed referendum day reminders to remind people to vote. These reached around 12.7 million users on Facebook with around 13.6 million impressions, and around 6.55 million users on Instagram with around 6.57 million impressions.[98] In 2022, our 'enrol to vote' prompt for the federal election was seen by 23.3 million Australians, with 54,000 people clicking through to the AEC website, and 900,000 people sharing it across their Feeds. The 'election day reminder' was seen by almost 11 million Australians, clicked through 175,000 times, and shared by 60,000 people on their Feed.[99]

We implemented a similar comprehensive strategy ahead of the 2022 Federal Election to combat mis- and disinformation on our services and promote civic participation, which we outlined in our 2023 transparency report under the Code. For example:

- During the campaign (between 1 April and 30 June 2022), we took action on:
    - over 25,000 pieces of content across Facebook and Instagram for violating our Harmful Health Misinformation policies.

---

[96] Meta, 'Meta Content Library and API', *Transparency Center*, https://transparency.meta.com/en-gb/researchtools/meta-content-library

[97] See eg., Meta Policy AU, 'How Meta is Preparing for the Voice to Parliament Referendum', Meta Australia Policy Blog, *Medium*, 10 July 2023, https://medium.com/meta-australia-policy-blog/how-meta-is-preparing-for-the-voice-to-parliament-referendum-282632baccfb

[98] Meta, 'Meta response to the Australian Code of Practice on Disinformation and Misinformation', May 2024, https://digi.org.au/wp-content/uploads/2024/05/Meta-Transparency-Report-2024-Australian-Code-of-Practice-on-Disinformation-and-Misinformation.pdf

[99] Meta, 'Meta response to the Australian Code of Practice on Disinformation and Misinformation', May 2024, https://digi.org.au/wp-content/uploads/2024/05/Meta-Transparency-Report-2024-Australian-Code-of-Practice-on-Disinformation-and-Misinformation.pdf

    - over 91,000 pieces of content on Facebook and over 40,000 pieces of content on Instagram in Australia for violating our hate speech policies.
    - over 200,000 pieces of content on Facebook and over 46,000 of content on Instagram in Australia for violating our Community Standards on violence and incitement.
- We displayed warnings on over 3 million distinct pieces of content on Facebook (including reshares) based on articles written by our third party fact checking partners.[100]
- We also rejected around 17,000 ads for not complying with our political and social issue ads enforcement policies.[101]

To deepen the understanding of the role of social media in democracy, we have also recently undertaken an unprecedented research partnership between Meta and external academics to better understand the impact of Facebook and Instagram on key political attitudes and behaviors during that election cycle.[102]

To date, four of 16 papers have been published and these initial four include studies of the effects of algorithmic ranking and virality, the prevalence and effects of like-minded information exposure on Facebook, and ideological segregation in exposure to news. Although questions about social media's impact on key political attitudes, beliefs, and behaviors are not fully settled, the experimental findings add to a growing body of research showing there is little evidence that key features of Meta's platforms alone cause harmful 'affective' polarisation or have meaningful effects on these outcomes. They also challenge the now commonplace assertion that the ability to reshare content on social media drives polarisation.

For example, Nature's summary of one paper states that the findings "challenge popular narratives blaming social media echo chambers for the problems of contemporary American democracy."[103] And the co-chairs of the study have stated: "Removing reshared content on Facebook produced a decrease in news knowledge among the study participants, and did not significantly affect political polarisation or other individual-level political attitudes."[104]

The studies also shed new light on the claim that the way content is surfaced on social media — and by Meta's algorithms specifically — keeps people divided. One of the papers shows there is considerable ideological segregation in consumption of political news, reflecting a complex

[100] Meta, Meta response to the Australian disinformation and misinformation industry code - reporting period January-December 2023 (published May 2023), https://digi.org.au/wp-content/uploads/2023/05/Meta_2023-AU-Misinformation-Transparency-report_v1.pdf
[101] Meta, Meta response to the Australian disinformation and misinformation industry code - reporting period January-December 2023 (published May 2023), https://digi.org.au/wp-content/uploads/2023/05/Meta_2023-AU-Misinformation-Transparency-report_v1.pdf
[102] Meta, 'Groundbreaking studies could help answer the thorniest questions about social media and democracy' (published July 2023), https://about.fb.com/news/2023/07/research-social-media-impact-elections/
[103] Nature, 'Like-minded sources on Facebook are prevalent but not polarizing' (published July 2023), https://www.nature.com/articles/s41586-023-06297-w
[104] Scimex, 'Expert reaction: Facebook and Instagram 'echo chambers' may not be driving our political polarisation' (published July 2023), https://www.scimex.org/newsfeed/facebook-and-instagram-echo-chambers-may-not-be-driving-our-political-polarisation

interaction between algorithmic and social factors. Yet, when participants in the experiments saw a reduced amount of content from sources that reinforced their views, they were actually more likely to engage with the like-minded content they did see. And even then, it had no detectable impact on polarisation, political attitudes or beliefs.

Scams

We adopt a four-pronged approach to combatting scams: (1) policies that prohibit scams and related behaviour; (2) enforcement both on and off-platform; (3) tools to allow people to block and report scams, but also warn people about potentially suspicious activity; and (4) consumer education initiatives and partnerships.

We recognise that the Australian Government expects industry to do more to combat scams, especially those targeting Australians. To respond to this expectation, we have stepped up our efforts to ingest more intelligence signals, deploy greater risk-based verification measures and increase our awareness initiatives in Australia.

We have a range of policies across our services that relate to scams, including:

- across Facebook and Instagram, we have a specific Fraud and Deception policy to protect people and businesses on our platform.[105] Under our Fraud and Deception policy, we remove content that purposefully deceives, willfully misrepresents or otherwise defrauds or exploits others for money or property. This includes content that seeks to coordinate or promote these activities using our services,
- our Advertising Standards strictly prohibit deception and misleading behavior,[106] and
- our Commerce Policies prohibit listings with misleading offers.[107]

On WhatsApp, we also have policies and systems that work to detect and enforce against abusive accounts, such as frauds and scams.[108]

In addition, we block the use of specific search terms related to scams, fake reviews, and known bait words. We also have measures in place to make Groups/Pages on Facebook that previously violated our policies less prominent in Feed and in recommendations. This is in line with our Content Distribution Guidelines, where we provide details about the problematic or low quality content that we reduce for distribution.[109] This includes content that contains clickbait links, engagement bait, links to websites that request unnecessary user data. We also exclude this type of content from being recommended across a range of surfaces, as outlined in our Recommendation Guidelines.[110]

---

[105] Meta, 'Facebook Community Standards - Fraud and Deception', *Transparency Center*, https://transparency.fb.com/policies/community-standards/fraud-deception
[106] Meta, 'Introduction to the Advertising Standards', *Transparency Center*, https://transparency.fb.com/policies/ad-standards
[107] Meta, Commerce policies, 'Prohibited content: Misleading, Violent, or Hateful', https://www.facebook.com/policies_center/commerce/misleading_violent_or_hateful
[108] WhatsApp, 'WhatsApp Business Messaging Policy', https://business.whatsapp.com/policy
[109] Meta, 'Types of Content We Demote', *Transparency Center*, https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote
[110] Facebook, 'What are recommendations on Facebook?', *Help Centre*, https://www.facebook.com/help/1257205004624246

Anti-scam partnerships and initiatives

Combatting scams is an ongoing challenge across many industries. When bad actors count on us to work in silos while they target people far and wide across the internet, we need to work together as an industry to protect people.

We have had a dedicated channel for ACCC to report scams content to us since September 2017. We review the content that is reported and take appropriate action if it is found to be violating.

Since the establishment of the National Anti-Scams Centre (NASC) in July 2023, we have actively engaged with the various processes that it has established and worked towards increased industry collaboration to identify what more all of industry can be doing to combat scams. Beyond removing individual scam reports, we are working closely with the NASC to also identify scams trends and address these.

In order to improve our detection, we make changes to our machine learning models by ingesting new scam trends and signals that we receive from users' reports and government escalations. We are also building a system for evaluating and ensuring the precision of our machine learning. Over time, these changes allow us to improve our proactive detection and enforcement at scale. We are also currently exploring the development of tools by introducing new technology that would allow our system to detect better, faster and receptive to the newer scams trends.

We are also working proactively with industry to develop a new Anti-Scams Code. This code is on track to be launched by the Digital Industry Group Inc. (DIGI) in the second half of 2024 and aims to uplift the efforts by digital platforms to combat scams in Australia.

To respond to the Government's call for industry to do more to combat scams targeting Australians, throughout 2024 we have been undertaking a suite of new measures:

- **New cross-industry escalation channels for Australian trusted partners**:
    - *The Fraud Intelligence Reciprocal Exchange Channel (FIRE)* – is a direct channel for trusted partners to report online fraud intelligence to Meta. To support streamlining reports from the banking industry in April 2024, we launched a new dedicated pilot channel for selected banks to fast track suspected scams on Facebook and Instagram directly to the relevant Meta teams. Information shared by this channel will help both parties to better reduce the harm generated by the ever-changing online scam landscape.
    - *New WhatsApp Trusted Partner Reporting Channel* – earlier in 2024, we launched a new WhatsApp Trusted Partner Reporting Channel for NASC to help streamline the reporting process.

- **New advertiser verification:** we are progressing plans to test lightweight verification for advertisers with heightened measures for higher risk areas, specifically:
    - *Higher levels of verification for new ad accounts* - from mid-June 2024, we began the global rollout of new advertiser verification, as part of which new advertisers may be required to have a verified phone number associated with their ad account before publishing ads. This involves an account administrator confirming a code sent by Meta through SMS, voice call or WhatsApp to validate their phone number. Our analysis indicates that the vast majority of ad scams detected originate from accounts that are less than 90 days old. We are therefore focusing our efforts on applying a higher level of verification to new ad accounts and test verification for new advertisers.
    - *Financial Service Ads Verification* – We are also looking to launch new ads verification and transparency measures for financial services ads.
- **New education resources and awareness campaigns including a launching a new Anti-Scams Resource Hub and new consumer campaigns with local partners**
    - *New anti-scams resource hub* - to expand our efforts to educate users and businesses on how to identify and avoid scams we launched a new anti-scams resource hub in the first half of 2024.[111] This hub includes information about the latest trends in scams, Meta's latest advances in cybersecurity, tips and educational material, and quick links for reporting scams, account access issues, and IP, brand rights protection and impersonation issues.
    - *New local scam awareness campaigns* – leveraging the new resource hub and our existing partnerships, we will launch new scam awareness and consumer campaigns in 2024, including new education materials addressing scam prevention on WhatsApp and cybersecurity for small businesses.

These initiatives build on our past efforts, including partnering with organisations such as IDCARE, Puppy Scam Awareness Australia and the Australian Small Business and Family Enterprise Ombudsman to deliver a scams awareness campaign in late 2021, including tips on how to identify different types of scams and report them, and account safety and cybersecurity tips. The campaign reached over 7.7 million people in Australia. In 2023, we worked with Australian creators to launch a new scam education campaign to coincide with 2023 national Scams Awareness Week, in partnership with local creator @joshandmattdesigns .

In late 2022, we updated our Meta Boost digital skills training curriculum for small businesses to include a new module on online safety and cybersecurity, which shared tips and advice for SMEs on how to protect their accounts and Pages from scams and fraudulent activity. We delivered this module as part of our Meta Boost training events in Byron Bay and Western Sydney in 2022 and 2023, in partnership with the Byron Bay, Mullumbimby and Majors Bay Chambers of

[111] Meta, *Anti-scam Hub*, https://about.meta.com/actions/safety/anti-scam

Commerce.[112]

Scam-related litigation

As part of Meta's ongoing efforts to enforce our Terms and protect people against abuse, we have brought legal action against individuals and entities responsible for using our platforms to scam people. For example:

- In 2019, we filed suit in California against a company called ILikeAd Media International Company Ltd. and two individuals for violating our Terms and Advertising Policies.[113]
- In 2021, we filed a case against four individuals residing in Vietnam, who used a technique known as "session theft" or "cookie theft" to compromise accounts of employees of advertising and marketing agencies and then ran unauthorised ads.[114]
- In 2022, Meta and a financial services company filed a joint lawsuit, the first of its kind, against two Nigerian-based individuals who engaged in phishing attacks to deceive people online and gain access to their online financial accounts. We had taken several prior enforcement actions against the defendants, including disabling Facebook and Instagram accounts, blocking impersonating domains on its services and sending a cease and desist letter. This joint lawsuit represented a major step forward in cross-industry collaboration against online impersonation.[115]
- In 2022, we filed a lawsuit against an Australian resident, Chad Taylor Cowan, for providing a fake engagement service directed at Facebook. Cowan operated a website that provided fake reviews and feedback to businesses in order to artificially increase their Customer Feedback Score.[116]

Content Distribution Framework

As there has been a growing amount of content shared online, it has been harder for people to find all of the content they cared about. This is why apps such as Facebook and Instagram use algorithms to connect people more quickly with content that they may find relevant.

We understand there is concern about the role of algorithms and AI in ranking and recommending content. This is why we prioritise providing greater transparency to help users better understand how our ranking algorithms and AI-powered products work and when they are engaging with AI-generated content, as well as provide users with more tools to control what they see in their Feed. Our President of Global Affairs, Nick Clegg, outlined Meta's

[112] See Meta Policy AU, 'Meta heads to Byron Bay to boost small businesses', *Medium*, 1 February 2023, https://medium.com/meta-australia-policy-blog/meta-heads-to-byron-bay-to-boost-small-businesses-da79e6a9574e; Meta Policy AU, 'Meta heads to Western Sydney to boost small businesses', *Medium*, 22 May 2023, https://medium.com/meta-australia-policy-blog/meta-heads-to-western-sydney-to-boost-small-businesses-4a2233d570c8
[113] Meta, 'Taking Action Against Ad Fraud', Newsroom, 5 December 2019, https://about.fb.com/news/2019/12/taking-action-against-ad-fraud
[114] Meta, 'Combating E-Commerce Scams and Account Takeover Attacks', *Newsroom*, 29 June 2021, https://about.fb.com/news/2021/06/combating-e-commerce-scams-and-account-takeover-attacks
[115] Meta, 'Taking Legal Action Against Financial Services Scams', *Newsroom*, 8 February 2022, https://about.fb.com/news/2022/02/taking-legal-action-against-financial-services-scams
[116] Meta, 'Taking Action Against Fake Customer Feedback and Reviews', *Newsroom*, 16 March 2022, https://about.fb.com/news/2022/03/taking-action-against-fake-customer-feedback-and-reviews/

approach on this in an article published last year.[117] People who use our products should have meaningful transparency and control around how data about them is collected and used, and this should be explained in a way that is understandable. That is why we are:

- Being meaningfully transparent about when and how AI systems are making decisions that impact the people who use our products;
- Informing people about the controls they have over those systems;
- Making sure these systems are explainable and interpretable; and
- Investing in research, explainability and collaboration.

At Meta, we use a range of different algorithms to help us rank content. The ones that people are often most familiar with are those that we use to rank content in their Feeds on Facebook and Instagram. Those algorithms that help with ranking play different roles. Some help us find and remove content from our platform that violates our Community Standards, or filter content that is potentially problematic or sensitive. Others help us understand what content is most meaningful to people so we can order it accordingly in their feeds. Below, we have outlined more information about these ranking algorithms as well as some of the algorithms we use to recommend new experiences to people.

It is important to bear in mind that the content people see in their Feeds is not solely due to algorithms: what people see is heavily influenced by their own choices and actions. Content ranking is a dynamic partnership between people and algorithms.Even though the people that use our services play a significant role in the ranking process, we recognise that they are only going to feel comfortable with these algorithmic systems if they have more visibility into how they work and then have the ability to exercise more informed control over them. That is why we have been releasing products, tools and greater transparency about the way algorithms work on our services. Our Content Distribution Guidelines[118] and Recommendation Guidelines,[119] explained in more detail below, both set a higher benchmark than our Community Standards; they apply to content that would not otherwise violate our rules on Facebook and Instagram.

Role of algorithms

“Algorithm” is a word that is often used but infrequently defined. In general, an algorithm is just a set of rules that help computers and other machine-learning models make decisions. Yet, in the context of social media, “algorithms” are often cited as a concern regarding the claimed influence of social media in promoting social polarisation and the spread of mis- and dis-information.

---

[117] Meta, ‘How AI Influences What You See on Facebook and Instagram’, *Newsroom*, 29 June 2023, https://about.fb.com/news/2023/06/how-ai-ranks-content-on-facebook-and-instagram/
[118] Meta, ‘Types of content we demote’, *Transparency Center*, 20 December 2021, https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/
[119] Facebook, ‘What are recommendations on Facebook?’, *Help Centre*, https://www.facebook.com/help/1257205004624246; Instagram, ‘What are recommendations on Instagram?’, *Help Centre*, https://help.instagram.com/313829416281232

Concerns regarding social media algorithms are also driven by a lack of understanding of the role of algorithms, and overlook the transparency and controls available to users to better understand and manage them.

On Facebook and Instagram, one of the ways that people connect with friends, family and other accounts that they follow is via a "Feed" .

Historically, these feeds showed content in chronological order, but as more people started using our services and more content was shared, it was impossible for people to see all of the content that was shared, much less the content that they cared about. Instagram, for example, launched in 2010 with a chronological feed but by 2016, people were missing 70 per cent of all their posts in Feed, including almost half of posts from their close connections. So we developed and introduced a Feed that ranked posts based on what people cared about most.[120] Similarly, on Facebook, the goal of Feed is to arrange the posts from friends, Groups and Pages people follow to show what matters most at the top of their feed. Our ranking algorithms use thousands of signals to rank posts for each person's Feed with this goal in mind.[121] As a result, each person's Feed is highly personalised and specific to them. Our ranking system personalises the content for over a billion people and aims to show each of them content we hope is most valuable and meaningful, every time they come to Facebook or Instagram.

Every piece of content that could potentially feature in a person's Feed — including the posts someone has not seen from their connections, the Pages they follow, and Groups they have joined, as well as content they could be interested in — goes through the ranking process. We call that universe of content someone's inventory. Because we have billions of people using our services and thousands of pieces of content that could potentially be seen in their Feed, we use the ranking process on trillions of posts across the platform.

From that initial inventory, thousands of signals are assessed for these posts, like who posted it, when, whether it is a photo, video or link, how popular it is on the platform, or the type of device you are using. In the next step from there, our ranking algorithms use these signals to predict how likely the post is to be relevant and meaningful to a person: for example, how likely a person might be to engage with it or find that viewing it was worth their time. The goal is to make sure people see what they will find most meaningful — not to keep people glued to their smartphone for hours on end.

One way we measure whether something creates long-term value for a person is to ask them. For example, we survey people[122] to ask how meaningful they found an interaction or whether a

[120] A Mosseri, 'Shedding more light on how Instagram works', *Instagram Blog,* 8 June 2021, https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works
[121] Meta, 'How does News Feed predict what you want to see?', *Newsroom,* 26 January 2021, https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see
[122] Meta, 'Using surveys to make News Feed more personal', *Newsroom,* 16 May 2019, https://about.fb.com/news/2019/05/more-personalized-experiences

post was worth their time, so that our system reflects what people enjoy and find meaningful.[123] Then we can take each prediction into account for a person based on what people tell us (via surveys) is worth their time.

While a post's engagement — or whether people like it, comment on it, or share it — can be a helpful indicator that it is interesting to people, this survey-driven approach, which largely occurs outside the immediate reaction to a post, gives a more complete picture of the types of posts people find most valuable, and what kind of content detracts from their Feed experience. We are continuously working on building out these surveys by asking new questions about the content people find valuable, and we have made it much easier for people to tell us what content they do not enjoy seeing in their Feed.[124]

In order to determine whether a post is likely to be valuable to people, our ranking process also assesses whether the post is likely to be problematic in some way. There are types of content and behaviour that do not violate our Community Standards, but users may tell us they do not like that form of content, so we use the ranking process to reduce their distribution. Other types of problematic content are addressed more directly through the ranking process. Some types of problematic content that receive reduced distribution through our ranking process include clickbait, unoriginal news stories, content likely to violate our Community Standards, and posts deemed false by one of the more than 90 independent fact checking organisations that review content on our apps. In 2021, we published a list of all of the types of problematic content and behaviour that receive reduced distribution on Feed, called our Content Distribution Guidelines, which we explain in more detail below.

After all of those steps, every post in a person's inventory receives what we call a "value score." In general, how likely a post is to be relevant and meaningful to people acts as a positive in the scoring process, and indicators that the post may be problematic (but non-violating) act as a negative. The posts with the highest scores after that are normally placed closest to the top of people's Feed.

Across our apps, we also make personalised recommendations to help users discover new communities and content we think they are likely to be interested in. Some examples of our recommendations experiences include Pages You May Like, "Suggested for You" posts in Feed, People You May Know or Groups You Should Join.

Since recommended content does not come from accounts that people have already chosen to follow, it is important that we have high standards for what we recommend. This helps ensure we do not recommend potentially sensitive content to those who do not explicitly indicate that they wish to see it. Our Recommendations Guidelines set a higher bar than our Community

[123] Meta, 'How users help shape Facebook', *Newsroom,* 13 July 2018, https://about.fb.com/news/2018/07/how-users-help-shape-facebook/
[124] Meta, 'Incorporating more feedback into News Feed ranking', *Newsroom,* 22 April 2021, https://about.fb.com/news/2021/04/incorporating-more-feedback-into-news-feed-ranking/

Standards, and content may be removed from recommendations even if it does not violate our Community Standards.

Policies

Providing guidelines for ranking
To increase the transparency around why people see particular content or ads, we provide transparency around ranking algorithms by publishing content ranking guidelines and details of any updates.

As mentioned above, we have published Facebook's Content Distribution Guidelines to share more detail on the types of content that we demote in Feed[125], and likewise for Instagram Feed and Stories. While the Community Standards make it clear what content is removed from our services because we do not allow it, the Content Distribution Guidelines make it clear what content receives reduced distribution because it is problematic or low quality. Many of these guidelines have been shared in various announcements, but in efforts to make them more accessible, we have brought them together in one easy-to-navigate space in our Transparency Center and Help Center.

The changes we make, particularly ones focused on limiting the spread of problematic content, are based on extensive feedback from our global community and external experts. Over the last few years, we have consulted more than 100 stakeholders across a range of relevant focus areas to solicit feedback on how to bring more insightful transparency to our efforts to reduce problematic content.

There are three principal reasons why we might reduce the distribution of content:

- **Responding to People's Direct Feedback.** We listen to people's feedback about what they like and do not like seeing and make changes to their Feeds in response.
- **Incentivising Creators to Invest in High-Quality and Accurate Content.** We want people to have interesting new material to engage with in the long term, so we're working to set incentives that encourage the creation of these types of content.
- **Fostering a Safer Community.** Some content may be problematic or sensitive for our community, regardless of the intent. We'll make this content more difficult for people to encounter.

Since 2021, we have also published a quarterly Widely Viewed Content Report (WVCR), which aims to provide more transparency and context about what people are seeing on Facebook by sharing the most-viewed domains, links, Pages and posts for a given quarter in Feed in the

[125] Meta, 'Types of content we demote', *Transparency Center,* 20 December 2021, https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/

United States.[126] The WVCR provides additional insights into the different content types that appear in News Feed to help people better understand our distribution systems and how that influences the content people see on our platform. We plan to expand the scope of this report to other countries in future iterations. It will continue to appear in conjunction with our quarterly Community Standards Enforcement Report.

We continually evaluate the effectiveness of Feed ranking signals. We are also making an effort to provide people with more detail about our ranking processes in general. For example, in 2021, the CEO of Instagram published a blog post detailing the ranking process on Instagram from start to finish, which we updated last year.[127]

Providing guidelines for recommendations

Across our apps, we make personalised recommendations to help users discover new communities and content we think they are likely to be interested in. Some examples of our recommendations experiences include Pages You May Like, "Suggested For You" posts in Feed, People You May Know or Groups You Should Join.

Since recommended content does not come from accounts that people have already chosen to follow, it is important that we have high standards for what we recommend. This helps ensure we don't recommend potentially sensitive content to those who don't explicitly indicate that they wish to see it. As noted above, our Recommendations Guidelines set a higher bar than our Community Standards, and content may be removed from recommendations even if it does not violate our Community Standards.

To help people better understand our approach to recommendations, in August 2020, we published a set of Recommendation Guidelines, which outline the types of content that may not be eligible for recommendations.[128] In developing these guidelines, we consulted 50 leading experts specialising in recommendation systems, expression, safety and digital rights. Recommendation Guidelines are available for both Facebook[129] and Instagram.[130]

---

[126] Meta, 'Widely Viewed Content Report: What People See on Facebook', *Transparency Center*, https://transparency.meta.com/en-gb/data/widely-viewed-content-report/
[127] A Mosseri, 'Shedding more light on how Instagram works', *Instagram Blog*, 8 June 2021, https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works; A Mosseri, 'Instagram Ranking Explained', *Instagram Blog*, 31 May 2023, https://about.instagram.com/blog/announcements/instagram-ranking-explained
[128] Meta, 'Recommendation guidelines', *Newsroom,* 31 August 2020, https://about.fb.com/news/2020/08/recommendation-guidelines/
[129] Facebook, 'What are recommendations on Facebook?', *Help Centre,* https://www.facebook.com/help/1257205004624246
[130] Instagram, 'What are recommendations on Instagram?', *Help Centre,* https://help.instagram.com/313829416281232

Transparency

Meta adopts a range of policies and transparency measures to help people be informed about the content they see on our services and why they are seeing it, and to protect them from harmful content such as dis- and mis-information.

We also recognise the important role that transparency and accountability play in giving policy makers, regulators and others insights into, and confidence in our investment and commitment to safety and integrity.

In addition to the transparency measures outlined above, we have a number of tools to provide people with greater insight and control over their experience. These include:

- *Why Am I Seeing this post?* - helps users to better understand and more easily control what they see from friends, Pages and Groups in their Feed. Users are able to tap on posts and ads in Feed, get context on why they are appearing (such as how their past interactions impact the ranking of posts in their Feed), and take action to further personalise what they see.[131] This includes the ability to customise their Feed, such as switching between an algorithmically-ranked Feed and a feed sorted chronologically with the newest posts first,[132] as well as indicating if they are interested or not interested in the post to inform future content recommendations.
- *Why Am I seeing this Ad?* - provides users with context on their ads, to help them understand how factors like basic demographic details, interests and website visits contribute to the ads in their Feed. We are continually improving our transparency offerings to reflect feedback we receive. In 2023, we updated this tool to provide users with clear information about the machine learning models that help determine the ads they see on Facebook and Instagram Feed.[133]
- *Ad Preferences* - allows users to adjust the ads they see while on Facebook and gives them the ability to update their ad settings to control information we can use to show their ads.[134]
- *Control what you see on Facebook and Instagram* - helps users to learn more about and control what kind of posts they may see on Facebook and Instagram, including who they see posts from.[135]

---

[131] Facebook, 'What influences the order of posts in your Facebook Feed', *Help Center*, https://www.facebook.com/help/520348825116417; Meta, 'Why Am I Seeing This? We Have an Answer for You', *Newsroom*, 31 March 2019, https://about.fb.com/news/2019/03/why-am-i-seeing-this

[132] Meta, 'More Control and Context in News Feed', *Newsroom*, https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/

[133] Meta, 'How does Facebook decide which ads to show me?', *Help Center*, https://www.facebook.com/help/562973647153813/?helpref=uf_share; Meta, 'Increasing Our Ads Transparency', *Newsroom*, https://about.fb.com/news/2023/02/increasing-our-ads-transparency, *Newsroom*, 14 February 2023

[134] Meta, 'Your Ad preferences and how you can adjust them on Facebook', *Help Center*, https://about.fb.com/news/2023/02/increasing-our-ads-transparency/

[135] Meta, 'Control what you see in Feed on Facebook', *Help Center*, https://www.facebook.com/help/1913802218945435/?helpref=uf_share; Instagram, 'How Instagram Feed Works', *Help Center*, https://help.instagram.com/1986234648360433

- *Content recommendation controls* - our content recommendation controls - known as "Sensitive Content Control" on Instagram and "Reduce" on Facebook – allows people to filter more potentially sensitive content or accounts from places like Search and Explore.[136]

As well as providing transparency at the user level, we recognise that there continue to be discussions about the best ways to provide model and systems documentation that enables meaningful transparency around how these systems are trained and operate. Our transparency initiatives at system level include the release of more than 22 AI System Cards that explain how the AI systems in our products work.[137] They give information, for example, about how our AI systems rank content, some of the predictions each system makes to determine what content might be most relevant to users, as well as the controls users can use to help customise their experience.

This is complemented by Meta's Transparency Center[138] and Privacy Center[139]. Our Transparency Center provides a one stop-shop that contains details of our policies, enforcement and integrity insights, such as:

- on our use of AI to inform ranking of content, our efforts to reduce problematic content and our AI-driven integrity efforts as part of our content governance;[140]
- our quarterly Community Standards Enforcement Report[141], which provides transparency about how we enforce our policies across Facebook and Instagram. This report provides data on how much harmful content we action, prevalence of harmful content, proactive detection rates as well as appealed and restored content;[142]
- our quarterly Adversarial Threat Report[143], where we regularly publish our findings about cyber threats we detect and remove across our technologies in our quarterly Adversarial Threat Report.[144] We do so to help our industry peers and security researchers better understand and counter internet-wide threats.

Our Privacy Centre informs people of how we build privacy into our products, including how we use information for generative AI models and features,[145] and how users can manage and control their privacy on Facebook, Instagram, Messenger and other Meta products. We also provide a

---

[136] Meta, 'Introducing Sensitive Content Control', *Newsroom*, 20 July 2021, https://about.fb.com/news/2021/07/introducing-sensitive-content-control; Facebook, 'Manage how content ranks in your Feed using Reduce', *Help Center*, https://www.facebook.com/help/543114717778091
[137] Meta Resources, System Cards, https://ai.meta.com/tools/system-cards/
[138] Meta, *Transparency Center*, https://transparency.meta.com/en-gb/
[139] Meta, *Privacy Center*, https://www.facebook.com/privacy/center
[140] Meta, 'Our approach to ranking explained', *Transparency Center*, June 2023, https://transparency.fb.com/features/explaining-ranking/
[141] Meta, 'Community Standards Enforcement Report', *Transparency Center*, https://transparency.fb.com/reports/community-standards-enforcement
[142] Meta, 'Community Standards Enforcement Report', *Transparency Center*, https://transparency.fb.com/data/community-standards-enforcement/
[143] Meta, 'Meta's Threat Disruptions', *Transparency Center*, https://transparency.meta.com/en-gb/metasecurity/threat-reporting/
[144] Meta, 'Meta's Threat Disruptions', *Transparency Center*, https://transparency.meta.com/en-gb/metasecurity/threat-reporting/
[145] Meta, 'How Meta uses information for generative AI models and features', *Privacy Center*, https://www.facebook.com/privacy/genai

deeper look at the types of signals and prediction models that we use in our ranking systems to reduce problematic content.[146]

Meta is a founding signatory of the DIGI Australian Disinformation and Misinformation Industry Code. Under this Code, Meta has committed to safeguards to protect people in Australia against harmful mis- and disinformation, and to adopting a range of scalable measures that reduce its spread and visibility. We have opted into all seven of the Code's objectives across Facebook and Instagram.

To date, Meta has published four transparency reports under the Code, with the latest launched in May 2024. Our 2024 report[147] outlines the steps we took during the 2023 calendar year to meet the 38 commitments we opted into over that reporting period, such as relating to global adversarial threats, our misinformation efforts in Australia, and our investment in media literacy and events. Meta's 2024 transparency report under the Code also shared a case study outlining our comprehensive strategy leading up to Australia's Aboriginal and Torres Strait Islander Voice Referendum in October 2023 to proactively detect and remove content that breached our services, combat misinformation, harmful content and, and promote civil participation.

[146] Meta, 'Our approach to Facebook Feed ranking', *Transparency Center*, June 2023, https://transparency.fb.com/en-gb/features/ranking-and-content/
[147] Meta, 'Meta response to the Australian Code of Practice on Disinformation and Misinformation', May 2024, https://digi.org.au/wp-content/uploads/2024/05/Meta-Transparency-Report-2024-Australian-Code-of-Practice-on-Disinformation-and-Misinformation.pdf

Industry & Regulatory Governance Frameworks

Industry partnerships

As noted above, we recognise the importance of industry working together to promote best practice, standard setting and collaboration which can work towards greater safety and security across the broader digital ecosystem. We have outlined above our industry partnerships such as Project Protect, the Tech Coalition, Lantern, and the GIFCT.

In addition, Meta is also a founding member of the Digital Trust and Safety Partnership (DTSP), which aims to set out global best practices to mitigate content and conduct-related risks associated with internet services and then verify that companies adhere to those best practices through internal and independent third-party assessments. It is a first-of-its-kind initiative bringing together technology companies of different sizes and business models around a common approach to increasing Trust & Safety across the internet. Members include Meta, Google, LinkedIn, Microsoft, Apple, Zoom, Pinterest, Reddit, Bitly, Discord, Twitch, TikTok, and the Match Group. In 2022, the DTSP published a new report entitled *"The Safe Assessments: An Inaugural Evaluation of Trust & Safety Best Practices"*[148] that synthesizes self-assessment submissions from ten DTSP members and provides an anonymized snapshot of industry's posture regarding digital trust and safety. Last year, the DTSP published a set of guiding principles and best practices to help industry define an overall age assurance framework.[149]

And finally, we proactively work with external stakeholders on hard questions in emerging areas, such as privacy, fairness, transparency, and safety in relation to AI. For example, we actively engage in multiple different fora and global conversations around appropriate AI governance frameworks, such as:

- Together with IBM, we established the AI Alliance,[150] 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.
    - Members and collaborators comprise companies, academic, and non-profit organisations that favour an open innovation approach to AI.[151] Notably, peer companies that advocate for a more proprietary approach to AI development - e.g., Google, Anthropic, OpenAI, and Microsoft - have not yet joined the Alliance.

[148] Digital Trust & Safety Partnership, 'DTSP Safe Assessments Report', https://dtspartnership.org/dtsp-safe-assessments-report/
[149] Digital Trust & Safety Partnership, *Age Assurance: Guiding Principles and Best Practices*, https://dtspartnership.org/wp-content/uploads/2023/09/DTSP_Age-Assurance-Best-Practices.pdf.
[150] AI Alliance, *Building the Open Future of AI*, https://thealliance.ai/
[151] See AI Alliance, *Meet the AI Alliance*, https://thealliance.ai/aia-members

- We are a founding member and funding support of the Partnership on AI (PAI),[152] a non-profit community of academic, civil society, industry, and media organisations creating solutions so that AI advances positive outcomes for people and society. We have representation on PAI's board, together with representatives from organisations such as Google Deepmind, Apple, Intel Labs, Microsoft, and OpenAI. We have been involved in several of PAI's initiatives, including:
    - Synthetic Media Framework.[153] A framework that provides specific recommendations on how to develop, create, and share content that has been generated or modified (usually via AI) in a responsible way. Recommendations focus on transparency, disclosure, and provenance measures, and some are tailored for different categories of stakeholder (e.g., those building the technology and infrastructure for synthetic media, those creating it, and those distributing and publishing it).
    - Guidance for Safe Foundation Model Deployment.[154] A framework that provides foundational model providers with a set of recommended practices to follow throughout the deployment process, tailored to the capabilities of their specific model (e.g., narrow or general purpose) and how it is being released (e.g., open source or behind an API).
- We are a founding member of ML Commons,[155] an AI engineering consortium, built on a philosophy of open collaboration to improve AI systems. Its mission is to create AI innovation and increase its positive impact on society by bringing together industry, academia, and non-profit organisations from around the world to develop open industry-standard benchmarks[156] and large-scale, diverse datasets[157] for model evaluation. We have representation on ML Commons' board, together with representatives from organisations including Google, Nvidia, Intel, Qualcomm, and Alibaba.
- We recently joined the Frontier Model Forum, a non-profit organisation and the only industry-supported body dedicated to advancing the safety of frontier AI models.[158]

We have participated in many of the international AI governance engagements, specifically, we supported the G7 Hiroshima AI Process, participated in the Bletchley UK AI Summit[159] and

[152] Partnership on AI, https://partnershiponai.org/
[153] Partnership on AI, 'PAI's Responsible Practices for Synthetic Media A Framework for Collective Action', https://syntheticmedia.partnershiponai.org/
[154] Partnership on AI, 'PAI's Guidance for Safe Foundation Model Deployment
A Framework for Collective Action', https://partnershiponai.org/modeldeployment/
[155] ML Commons, 'Better AI for Everyone', https://mlcommons.org/
[156] ML Commons, 'Benchmarks', https://mlcommons.org/benchmarks/
[157] ML Commons, 'Datasets', https://mlcommons.org/datasets/
[158] See Frontier Model Forum, 'Amazon and Meta join the Frontier Model Forum to promote AI safety', 20 May 2024, https://www.frontiermodelforum.org/updates/amazon-and-meta-join-the-frontier-model-forum-to-promote-ai-safety/
[159] UK Government, 'Chair's Summary of the AI Safety Summit 2023, Bletchley Park', https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park

signed the Seoul Frontier AI Safety Commitments[160]. As AI Safety Institutes[161] are established and commence their work in the US, UK and Japan, among others, we support future work plans to undertake effective evaluation of models. We have also publicly committed to Child Safety Generative AI Principles developed by Thorn and All Tech Is Human.[162]

In February 2024, we also signed on to *A Tech Accord to Combat Deceptive Use of AI in 2024 Elections*, pledging to work with other technology companies including Amazon, Google, and OpenAI, to help prevent deceptive AI content from interfering with this year's global elections.[163] This work is bigger than any one company and will require a huge effort across industry, government, and civil society.

Oversight Board

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.[164]

The Oversight Board was borne out of the recognition that critical decisions about content should not be left to companies alone. Content decisions can have significant consequences for free expression and companies like Meta - notwithstanding our significant investments in detection, enforcement and careful policy development - will not always get it right.

The Oversight Board comprises 22 experts in human rights and technology, including the Australian academic Professor Nicholas Suzor from Queensland University of Technology.[165] The Board is entirely independent and hears appeals on Meta's decisions relating to content on Facebook, Instagram and Threads. We have agreed that the Board's decisions will be binding, and the Board is also able to make recommendations about Meta's policies.[166]

---

[160] See UK Government, 'Historic first as companies spanning North America, Asia, Europe and Middle East agree safety commitments on development of AI', *Press Release*, 21 May 2024, https://www.gov.uk/government/news/historic-first-as-companies-spanning-north-america-asia-europe-and-middle-east-agree-safety-commitments-on-development-of-ai

[161] In 2023, the UK Government established an AI Safety Institute to equip governments with an empirical understanding of the safety of advanced AI systems (https://www.aisi.gov.uk/). Earlier this year, both the US Government and Japanese Governments also launched AI Safety Institutes (https://www.nist.gov/aisi and https://aisi.go.jp/ respectively).

[162] See Thorn, 'Thorn and All Tech Is Human Forge Generative AI Principles with AI Leaders to Enact Strong Child Safety Commitments', 23 April 2024, https://www.thorn.org/blog/generative-ai-principles

[163] AI Elections Accord, *A Tech Accord to Combat Deceptive Use of AI in 2024 Elections*, https://www.aielectionsaccord.com/

[164] Meta, 'Oversight Board to Start Hearing Cases', *Newsroom*, 22 October 2020, https://about.fb.com/news/2020/10/oversight-board-to-start-hearing-cases

[165] Oversight Board, *Get to know our Board members*, https://www.oversightboard.com/meet-the-board/

[166] Meta, 'Establishing structure and governance for an independent oversight board', *Newsroom*, 17 September 2019, https://about.fb.com/news/2019/09/oversight-board-structure/; Oversight Board, 'Providing an independent check on Meta's content moderation', https://www.oversightboard.com

The Oversight Board began issuing decisions in January 2021[167] and also policy advisory opinions. This included a decision and policy recommendation related to our COVID-19 misinformation and harm policies.

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.[168]

We believe the Oversight Board is a significant innovation in content governance and a first-of-its-kind initiative. It makes Meta more accountable for our content decisions and helps to improve our decision-making.

In 2024, the Oversight Board investigated two cases relating to the Australian Electoral Commission's (AEC) voting rules around the 2023 Aboriginal and Torres Strait Islander Voice Referendum.[169] The cases involved two separate Facebook posts containing the same screenshot of information posted on X by the AEC ahead of the Referendum. The information shown included the message that: 'If someone votes at two different polling places within their electorate, and places their formal vote in the ballot box at each polling place, their vote is counted.' The posts were removed by Meta for violating the rule in our Coordinating Harm and Promoting Crime Community Standard that prohibits content calling for illegal participation in a voting process. Both users had appealed the removal of these posts.

On 9 May 2024, the Oversight Board published its decision on the two cases, in which it upheld Meta's decision to remove the posts, as the two users' 'calls for others to engage in illegal behaviour impacted the political rights of people living in Australia' and that Meta 'was correct to protect democratic processes by preventing voter fraud attempts from circulating on its platforms, given the frequent claims that the Voice Referendum was rigged'.[170]

Australian regulatory compliance

We recognise the importance of greater transparency, accountability and user empowerment that is required for smart regulatory frameworks. This is why we have long been calling[171] for

[167] N Clegg, 'Welcome the oversight board', *Newsroom,* 6 May 2020, https://about.fb.com/news/2020/05/welcoming-the-oversight-board
[168] See Oversight Board, https://transparency.fb.com/en-gb/oversight
[169] Oversight Board, 'Oversight Board announces cases involving the Australian Electoral Commission's voting rules', February 2024, https://www.oversightboard.com/news/303530418986569-oversight-board-announces-cases-involving-the-australian-electoral-commission-s-voting-rules
[170] Oversight Board, 'Oversight Board upholds Meta's decisions in Australian Electoral Commission voting rules cases', May 2024, https://www.oversightboard.com/news/oversight-board-upholds-metas-decisions-in-australian-electoral-commission-voting-rules-cases
[171] Meta, 'Four Ideas to Regulate the Internet', 30 March 2019, https://about.fb.com/news/2019/03/four-ideas-regulate-internet/

government regulation for digital platforms, working to establish proactive regulatory models and contributing constructively to the debate surrounding digital policy. Given the many new laws that have been enacted or proposed in Australia specifically focused on digital platforms, the question is no longer whether regulation is needed, but whether it is effective at driving appropriate investment in safety and security across the industry. We welcome the opportunity to contribute to the debate around the efficacy of these regulatory frameworks.

In particular, the impact of the recent safety reforms should be considered. Since the new Online Safety Act took effect in 2022, we have worked to be a constructive industry contributor to the development of the six industry codes and two industry standards, and will continue to do so in respect of the new codes being worked on by industry this year. We have also responded to two rounds of transparency notices under the Basic Online Safety Expectations (BOSE) regime. This transparency under the BOSE regime comes in addition to our existing online safety investments and broader transparency and accountability. Finally, we have dedicated internal teams who have spent the last year coordinating our response to the new obligations under the codes and the broader online safety regulatory frameworks in Australia and globally. However, as part of the Statutory Review of the Online Safety Act, it is timely to consider the efficacy of this framework and how to benchmark industry compliance. Given the ongoing state of reform of Australia's online safety laws, there is a risk that they incentivise industry to invest in compliance systems rather than safety and security mitigations.

We take our legal obligations around the safety of users seriously and will continue to work constructively with policymakers and regulators on the development and implementation of new and existing laws.

Changes in news on Facebook

As outlined above, the way in which people have used Meta's services and the product innovations in which we have invested have changed over the years. Changes in user behaviour, technology shifts and the response to our news product investments, have meant that there are changes in how we invest in commercial deals with Australian publishers.

News content and investment in the Australian news

Meta has invested and provides ongoing value to Australian publishers in three ways – firstly, we invest in the infrastructure, product and integrity systems that provide Australian publishers with free distribution to connect with new audiences and commercialise this in the manner of their choosing; secondly, we have built customised products and supported these with commercial deals; and finally, we have invested in funds and programmatic initiatives to support innovation by diverse publishers.

With respect to commercial deals, we entered into commercial deals between 2021 to support release of the Facebook News product in Australia in August 2021.[172] In February 2024, we announced that we were deprecating the Facebook News product as part of our ongoing efforts to better align our investments to our products and services people value the most, including short form video.[173] The number of people using Facebook News in Australia and the U.S. dropped by over 80% last year and we're focusing our time and resources on things people tell us they want to see more of on the platform. We are continuing to honour the terms of the commercial deals that supported this product. These deals, and associated payments, started to expire at the end of May 2024, with the final one ending in December 2024.

In addition to agreements to support the Facebook News product, we also entered into agreements at the same time to support the Facebook Video product.[174] These deals expire between March and December 2024.

Recognising that not all publishers have the capacity to enter into commercial agreements and meet relevant performance requirements, we also launched two journalism funds to support innovation, public interest journalism and smaller publishers who would not be suitable for commercial agreements.

---

[172] Meta, 'Facebook Announces AU$15 million news fund and begins the phased launch of Facebook News in Australia', 4 August 2021, *Meta for Media blog*, https://www.facebook.com/journalismproject/facebook-invests-in-australia-news-fund-and-launches-facebook-news

[173] Meta, 'An update on Facebook News', *Newsroom*, 29 February 2024, https://about.fb.com/news/2024/02/update-on-facebook-news-us-australia/

[174] Meta, 'Facebook partners with Australian news publishers to fund news shows on Watch', *Meta for Media blog*, 5 August 2019, https://www.facebook.com/formedia/blog/facebook-partners-with-australian-news-publishers-to-fund-news-shows-on-facebook-watch

The first is a three-year innovation fund – the Newsroom Sustainability and Digital Transformation Fund – with Country Press Australia to support regional news that commenced in 2022[175]. In the first tranche, 106 regional mastheads across Australia were announced as recipients of funding. In the second tranche, funding was distributed in 2023 to 118 mastheads and, in 2024 funding went to 137 newsrooms.

The second fund – the Meta Australian News Fund – was established in 2021 via a three-year partnership with the Walkley Foundation to fund regional newsrooms, digital-first publications and public interest journalism projects with a total of AU$15 million provided over three tranches. The first tranche of AU$5 million was paid in 2022 towards 54 projects.[176] The second tranche of AU$5 million was paid in 2023 towards 45 projects. The third and final tranche of AU$5 million was announced in February 2024 and paid to 51 publishers.[177]

Since these commercial agreements and funds were established, there has been a change in consumer behaviour on our services. As a general rule, most people do not come to our services for news and news is highly substitutable on our services – this means that when news is not on our services, people continue to engage with other content. However, even allowing for this, since 2021, people have engaged increasingly with more non-news content, specifically short form video. The realignment of our investments is part of an ongoing effort to better align our investments to our products and services people value the most. However, whilst the investment of these commercial deals and funds will cease at the end of 2024, Meta continues to provide services that deliver value to Australian publishers through the free distribution that is available when they choose to share content on our services.

News organisations choose to share their content on Facebook and Instagram. By taking advantage of this free distribution news businesses can grow their audiences, sell subscriptions and boost ad revenue. Publishers keep 100% of the revenue from traffic and subscriptions derived from outbound links on both services. For example, in 2023 we estimate that Facebook Feed sent Australian publishers more than 2.3 billion free clicks — for no charge — driving an estimated (AUD) $115 million worth of value.[178]

Meta's investment in integrity systems

Meta invests significantly in technology, people and systems to combat coordinated inauthentic behaviour that can lead to misinformation and disinformation on our services. As part of this

[175] Meta, '106 Publishers Awarded Meta Country Press Association Newsroom Fund', *Meta for Media blog*, 11 April 2022, https://www.facebook.com/formedia/blog/meta-announces-recipients-of-country-press-australia-news-fund
[176] The Walkley Foundation, *Meta Australian News Fund*, https://www.walkleys.com/valuing-journalism/meta-australian-news-fund/
[177] The Walkley Foundation, 'Walkley Foundation announces 51 newsrooms and independent journalists to share AU$5M in Meta Australian News Fund third round', 8February 2024, https://www.walkleys.com/walkley-foundation-announces-51-newsrooms-and-independent-journalists-to-share-au5m-in-meta-australian-news-funds-third-round/
[178] Meta, 'Debunking claims about news content on Meta's platforms', Meta Australia Policy Blog, 13 March 2024, https://medium.com/meta-australia-policy-blog/debunking-claims-about-news-content-on-metas-platforms-b7117945ac87

investment, our technology reduces the distribution of content that is problematic or low-quality. This reduction is in response to direct feedback from people using our services. It incentivises publishers to invest in high-quality content, and fosters a safer community. We want people to be able to enjoy and share content without being disrupted by problematic or low-quality content.

Consequently, we reduce the distribution of unoriginal news articles. This includes news articles that don't contain new, original reporting or analysis. The more extensive original reporting an article contains, the more distribution it will receive in Feed, relative to less original reporting. Original reporting includes things such as exclusive source materials, significant analysis, new interviews or the creation of original visuals.

Additionally, Meta partners with third-party fact-checking organisations (**3PFC**), globally and in Australia, to assess the accuracy of content on our services. We have commercial arrangements with independent 3PFC organisations for them to review and rate the accuracy of posts on Facebook and Instagram. In Australia, we partner with Australian Associated Press, Agence France Presse and RMIT FactLab, all of which are certified by the non-partisan International Fact-Checking Network, as part of a network of over 90 fact-checking partners around the world covering more than 60 languages. All fact-checks by these partners are publicly available on their websites.[179] We outline in more detail above about how our fact-checking partnerships operate as part of our broader content integrity efforts to combat dis- and mis-information.

Changes in news consumption habits

The way in which people interact with technology changes over time, and digital platforms such as Meta must invest and innovate vigorously to compete and respond to these changes. We have consistently provided data and evidence to the Australian Government about the economics of the value of news on our services, that it is substitutable, and the changes in consumer behaviour with respect to news on our services. We have also provided evidence that there is no evidence supporting the assertion that there has been an increase in misinformation in Canada when news has been removed. Specifically:

- Evidence of the economic value of news on our services & changing news consumption habits:
    - In March 2023, NERA Economic Consulting published a report, commissioned by Meta, on “Meta and the News: Assessing the Value of the Bargain”[180] which analysed survey data from the Reuters Institute for the Study of Journalism of

[179] Agence France Presse Australia, Fact Check, https://factcheck.afp.com/afp-australia; Australian Associated Presse, AAP Fact Check, https://www.aap.com.au/category/factcheck/; RMIT FactLab, https://www.rmit.edu.au/about/schools-colleges/media-and-communication/industry/factlab/debunking-misinformation

[180] JA Eisenach, ‘Meta and the News: Assessing the Value of the Bargain’ Nera Economic Consulting, March 2023 https://www.nera.com/content/dam/nera/publications/2023/Meta%20and%20the%20News%20Assessing%20the%20Value%20of%20the%20Bargain.pdf

people across 26 countries who reported using Facebook from 2016 to 2022 for "finding, reading, watching, sharing or discussing news". The findings in this report counteract many claims being made by Australian news publishers about the value exchange that occurs using our services. It finds that consumption habits for news are changing (as noted further below) and news is not a substantial part (and is in fact a declining part) of Facebook (for example, 20% of Australians are reported seeing "too much" content from news outlets on Facebook).[181]

- The amount of referral traffic to Australian news publishers from Facebook Feed has declined over time. We have at various points reported figures publicly which show this. For example, we have reported that there were approximately 5.1 billion organic referrals or clicks in 2020 from Facebook Feed to Australian news publishers,[182] which declined to more than 3.5 billion in the 12 months to March 2022,[183] which declined again to more than 2.3 billion in 2023 due to the continuing shift in user preferences.[184]
- Similarly, engagement with Facebook News has declined dramatically over the time it was available. For example, the number of daily active users of Facebook News in Australia dropped over 80% in 2023. The same was the case in the US.

- <u>Evidence that news is substitutable:</u>
  - We have observed no meaningful impact to user engagement following the restrictions on the viewing and sharing of news content in Canada. Just as the number of people around the world using our technologies continues to grow, the number of Daily Active Users and Monthly Active Users on Facebook in Canada has increased since ending news availability. In addition, time spent on Facebook in Canada has continued to grow since ending news availability. Given Instagram is a visual-first service and was always significantly less of a destination for news content than Facebook, we would expect there to have been no meaningful impact on Instagram. We have not identified any other notable impacts to our business in Canada or around the world as a result of ending news availability in Canada.
  - These findings are consistent with public reporting. For instance, independent analysis by Reuters has also found that our decision to end news availability in Canada "had almost no impact on Canadians' usage of Facebook, data from independent tracking firms indicated" and "daily active users of Facebook and

[181] See Reuters Institute / Oxford University, 'Digital News Report 2022', pg 26, https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022
[182] Meta, 'The Real Story of What Happened With News on Facebook in Australia', 24 February 2021, https://about.fb.com/news/2021/02/the-real-story-of-what-happened-with-news-on-facebook-in-australia/
[183] Meta, 'Meta's response to the Treasury Department's review of the news media bargaining law', May 2022, https://treasury.gov.au/sites/default/files/2023-02/c2022-264356-meta.pdf
[184] Medium, 'Debunking claims about news content on Meta's platforms', 13 March 2024, https://medium.com/meta-australia-policy-blog/debunking-claims-about-news-content-on-metas-platforms-b7117945ac87

time spent on the app in Canada stayed roughly unchanged" based on data shared by Similarweb.[185]

- Publishers have many channels to engage with their audiences:
    - News is not uniquely valuable to our platform. In that regard, we note:
        - Digital audiences have never been better served through the multiple channels today for news content, and publishers have never had as much choice in reaching their audiences. From publishers' own websites and apps, to newsletters, podcasts, and digital video distribution channels (and platforms providing infrastructure to support those like Substack), to search services like Google search and Bing, news aggregators like Apple News and Google News, other platforms like LinkedIn, X, YouTube, TikTok, as well as outdoor screens, we represent but a few of more than a dozen digital touchpoints that are now available for publishers to reach audiences. For example, prior to ending news availability in Canada, one of Canada's major publishers, CBC, informed its audience of the 12 different sources for CBC news available to them in addition to Google, and (at the time) Facebook and Instagram.[186] In Australia, the breadth of different channels is likewise evident, for example, SBS News' site promotes its multi-channel delivery which spans 14 different touchpoints for consumers, only one of which is following SBS News on social channels which are not limited to Facebook and Instagram.[187]
        - Publishers in Australia also have that range of choices available and have been shifting into or expanding in additional channels, likewise showing that our services - while they may be popular in their use today - are by no means essential to publishers. For example, the Australian Financial Review reported the founder of Broadsheet, Nick Shelton (a registered news publisher on the ACMA's register of eligible news businesses), as saying he was *"not concerned with"* diminishing traffic from Facebook: *"Facebook once supplied 50 per cent of the culture publisher's traffic. That figure is now about 20 per cent, and declining rapidly. It's not something we're particularly concerned with. Our audience distribution focuses have long been on other things, namely owned audiences and direct relationships,"* Mr Shelton said. *"I would anticipate that publishers have been anticipating this move and have shifted into other channels, regardless. We build an app, we're building newsletters, and we're producing content we believe our readers are actively coming back for."*[188]

[185] Reuters, 'Exclusive: Meta's Canada news ban fails to dent Facebook usage', 29 August 2023, https://www.reuters.com/technology/metas-canada-news-ban-fails-dent-facebook-usage-2023-08-29
[186] CBC, 'Canadian news is starting to vanish from Instagram. Google is next. Here's how to find CBC as that happens', 5 July 2023, https://www.cbc.ca/news/editorsblog/cbc-online-news-act-1.6897060
[187] See SBS News, https://www.sbs.com.au/news
[188] Australian Financial Review, 'Facebook traffic to news publishers has plummeted 50pc this year; 10 September 10 2023, https://www.afr.com/companies/media-and-marketing/facebook-traffic-to-news-publishers-has-plummeted-50pc-this-year-20230908-p5e34p

- The Australian Financial Review also reported Misha Ketchell, editor of The Conversation (also a registered news publisher on the ACMA's register of eligible news businesses), as saying that *"Facebook is currently the source for only 4 percent of the publication's Australian traffic - a figure that has halved since January [2023]. "There has been a really steep decline. We don't know why. We suspect and have seen a really waning interest in news from Facebook,"* he said." [189]
- We also note that publishers have many channels with which to engage with consumers. For example, News Corp reported in its Q3 FY2024 earnings call an increase of 7% of digital subscribers at News Corp Australia offsetting lower digital advertising revenues "driven by a decline in traffic at some mastheads due to platform related changes",[190] reflecting a "conscious strategic shift away from potentially volatile advertising revenues to growth in circulation and subscription revenues".[191]

- <u>No evidence of increased misinformation when news is removed:</u>
  - We are aware of publisher allegations that ending news availability in Canada (or more generally reducing exposure to news content on our services) has led to an increase in misinformation/disinformation on our services. This is inaccurate, and we are not aware of any evidence supporting this assertion.
  - We have never thought about news as a way to minimize misinformation/disinformation on our services. With or without news content, we are incentivised to – and do (as our submission outlines above) – remove harmful misinformation and reduce distribution for fact-checked misinformation, and we remain steadfast in our commitments to ensure the integrity of information on our platforms by countering this type of harmful content. Canadians can continue to use our services to access authoritative information from a range of sources, including government agencies, political parties and non-governmental organisations, which have always shared information with their audiences in engaging formats, in addition to links to news content. Our fact-checking efforts apply to the type of content that remains available to people in Canada, and we are committed to stopping the spread of misinformation on our services.
  - We also note that in Australia, we are founding signatories of The Australian Code of Practice on Disinformation and Misinformation, as part of which we commit to undertake regular digital literacy initiatives, research on misinformation and annual transparency reports. This will continue to provide Australian policy

[189] Australian Financial Review, 'Facebook traffic to news publishers has plummeted 50pc this year; 10 September 10 2023, https://www.afr.com/companies/media-and-marketing/facebook-traffic-to-news-publishers-has-plummeted-50pc-this-year-20230908-p5e34p

[190] News Corp, 'News Corporation reports third quarter results for fiscal 2024', https://newscorp.com/wp-content/uploads/2024/05/Q3-FY2024-Earnings-Release_FINAL_8-May-2024.pdf

[191] Seeking Alpha, 'News Corporation (NWSA) Q2 2024 Earnings Call Transcript', 7 February 2024, https://seekingalpha.com/article/4668413-news-corporation-nwsa-q2-2024-earnings-call-transcript?source=content_type%3Areact%7Csection%3AAll%7Csection_asset%3ATranscripts%7Cfirst_level_url%3Asymbol%7Cbutton%3ATitle%7Clock_status%3ANo%7Cline%3A2

makers and other stakeholders accountability and transparency over Meta's ongoing work to combat mis- and disinformation.