



Meta's Submission to the Australian Select Committee on Social Media and Online Safety

JANUARY 2022

Executive summary

Meta welcomes the opportunity to provide a submission to the House Select Committee on Social Media and Online Safety.

We recognise our responsibility to protect the safety of people who use Meta's services – especially the safety of young people. It's essential to our business: Australians and other people around the world will only continue to use our platform if they feel welcome and safe.

Industry, government and the community all have a role to play in working towards online safety. To uphold our responsibility, we have made significant investments in our ability to keep people safe. At a global level, we now have more than 40,000 people working on safety and security at Meta. We've invested more than US\$13 billion (~AU\$18 billion) on safety and security since 2016, and we spent more than US\$5 billion (~AU\$6.9 billion) in 2021.

This submission provides a detailed overview of the various policies, enforcement techniques, tools, products, resources and partnerships that we have developed to protect the safety and security of our users. We have listed the initiatives we undertake to protect online safety (including of young people), combat misinformation, take action on hate speech, protect privacy and security, and encourage positive and healthy use of our services.

Our efforts are having an impact and we are making progress.

In our Community Standards Enforcement Report, we report every quarter against objective metrics on our efforts to combat harmful content.¹ We contend that there are two key ways of assessing the effectiveness of our systems and processes: how prevalent harmful content is for our users; and what proportion of harmful content removed is detected by us, before it's reported.

On both fronts, we've made considerable progress over the last few years. For example, we've cut the prevalence of hate speech content by more than half within the last year alone (from 0.07 to 0.08 per cent in Q3 2020, down to less than 0.03 per cent in Q3 2021). Moreover, for many categories of seriously harmful content, we are proactively detecting more than 99 per cent of content ourselves, before a user needs to report it to

¹ Meta, *Community Standards Enforcement Report* <https://transparency.fb.com/data/community-standards-enforcement/>

us. This is the case for content such as child exploitation material, terrorist content, violent and graphic content, and fake accounts.

Our tools to promote positive engagement are also working. For example, we've recently launched a tool on Instagram that sends a user a warning and discouragement when they draft a comment that is similar to bullying and harassment comments. Our experience to date is that, about 50 per cent of the time, the comment was edited or deleted by the user based on these warnings. Similarly, our efforts to promote authoritative COVID-19 information to combat misinformation has reached a large number of Australians: more than 6 million Australians accessed our COVID-19 Information Centre in 2020.

We've also taken steps to encourage accountability and oversight of our content decisions. We established an Oversight Board to make binding rulings on difficult and significant decisions about content on Facebook and Instagram. 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 40 experts in human rights and technology, including the Australian academic Professor Nic Suzor from Queensland University of Technology.

Concern has been expressed by some commentators that social media fuels polarisation, exploits human weaknesses and insecurities, or creates echo chambers. The factor that is often cited is concern around "algorithms". In addition to the work we've done to combat harmful content, we've also taken steps in order to be more transparent about how algorithms work, and to give people greater control. For example: in 2021, we released the industry-leading Content Distribution Guidelines², which outline the types of content that do not violate our Community Standards but will receive reduced distribution on News Feed because it's problematic or low quality. We have released a number of tools in order to give users greater transparency (such as Why Am I Seeing This?) and greater control (such as News Feed Preferences) over algorithms.³

Above and beyond the actions we've taken, we appreciate the important role that regulation plays in giving policymakers and the broader community confidence about safety and social media services. Meta has been at the global forefront of calling for new

² Meta, Types of content we demote, *Transparency Centre*, 20 December 2021, <https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/>

³ For more details, please see the "Algorithms" section below.

regulation, especially in areas such as content and online safety, privacy, elections and data portability since 2019.⁴

Consistent with this global commitment, we have also supported and encouraged regulation in Australia. For example: we were the first company to publicly endorse the eSafety Commissioner's Safety by Design Guidelines;⁵ we funded expert research on best practice misinformation regulation by an Australian academic in February 2021;⁶ and we were a critical driver in landing a world-leading industry code on misinformation and disinformation in 2021.⁷

The Committee should recognise that the picture of regulation of digital platforms in 2022 looks very different to 5 or 10 years ago. Indeed, the often-repeated allegations that social media is unregulated is completely false. Meta is highly responsive to Australian regulators and regularly restricts access to content on our services to respect Australian law.

The Australian Government has been among the most active in the world in introducing new regulations specifically related to digital platforms. In the last three years, at least 14 new federal regulations have come into force which primarily impact digital platforms. There have also been at least 18 major Government or parliamentary inquiries or consultations impacting digital platforms over the last three years. These developments are on top of existing regulations that cover digital platforms, including online safety, privacy and multinational taxation laws. The Government has recently foreshadowed additional regulations, including digital platforms-specific competition laws; age or identity verification; and new obligations about working with law enforcement.

Meta has responded constructively to all these inquiries, and we have supported many of the new laws that have resulted from them (including, principally, the Online Safety Act). Given the recent history of active rule-making, we suggest therefore that this Committee should focus its attention on whether the slew of regulations are effective or necessary.

⁴ M Zuckerberg, The Internet Needs New Rules, *Washington Post*, 30 March 2019, https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html

⁵ Safety by Design Youth Jam, *Facebook*, August 2019, <https://www.facebook.com/MetaAustralia/videos/910843179301219/>

⁶ A Carson, 'Fighting Fake News: A Study of Online Misinformation Regulation in the Asia-Pacific', https://www.latrobe.edu.au/_data/assets/pdf_file/0019/1203553/carson-fake-news.pdf

⁷ For more details, please see the "Misinformation" section below.

Policymakers should be alive to the risk of overlapping, duplicative or inconsistent rules across different laws. Indeed, many of the online safety-related laws and regulations that have already been passed by Parliament are yet to be implemented. Policymakers will be able to develop more effective regulation if there is consideration given to properly understanding the effectiveness of existing regulation first. We make recommendations to assist in this regard: we also recommend that the Committee address what ‘success’ looks like in relation to online harmful content. A review of the effectiveness of existing digital platforms regulation would be timely in 2023.

Additionally, the overall regulatory approach taken by Australia needs to be viewed in the context of a global contest of competing visions of the internet. Other countries look to Australia, and it is important to consider whether Australian regulation sets an example which encourages a liberal, open and democratic approach to the internet, or an internet that is more closed, tightly controlled and fragmented. Given threats to the integrity of the global internet, including data localisation laws, we encourage the Australian Government to work with other democratic governments and international organisations on broad agreement around internet regulation.

Meta will continue to be a constructive partner for Australian policymakers in considering these policy questions, and the best way to approach them, and welcomes the opportunity to engage with this inquiry.

Table of contents

| | |
|---|----|
| Executive summary | 2 |
| Table of contents | 6 |
| Safety | 7 |
| Supporting young people and parents | 15 |
| Bullying and harassment | 16 |
| Women’s safety | 21 |
| Public figures | 26 |
| Ensuring age-appropriate experiences online | 30 |
| Mental health and wellbeing | 36 |
| Hate speech | 43 |
| Misinformation | 47 |
| Transparency and accountability | 71 |
| Transparency Reports | 72 |
| Independent Oversight | 76 |
| Privacy and data security | 78 |
| Internal data governance | 78 |
| Privacy in our products | 79 |
| Privacy tools | 79 |
| Privacy enhancing technologies | 82 |
| Data security | 83 |
| Current and future regulation | 88 |
| Current regulatory landscape | 91 |
| Regulatory inconsistency | 91 |
| Combatting fragmentation of the internet | 94 |
| Summary of recommendations | 94 |

Safety

We recognise our responsibility to protect the safety of people who use Meta's services – especially the safety of young people. It's essential to our business: Australians and other people around the world will only continue to use our platform if they feel welcome and safe.

Meta makes significant, industry-leading investments to protect the safety of the community on our platform. We now have more than 40,000 people working on safety and security across the company, and we've invested more than US\$13 billion (~AU\$18 billion) in safety and security since 2016. In 2021 alone, we spent more than US\$5 billion (~AU\$6.9 billion) on safety and security.

This investment is focused on our industry-leading program of online safety that comprises five components. These components are each explained in detail below.

1. Policies
2. Enforcement
3. Tools and products
4. Resources
5. Partnerships.

Before turning to online safety measures specific to sections of the community that have been the focus of recent public debate – young people and parents, women and public figures – the Committee may find it helpful to have more information about our general approach to safety across all our users.

Policies

First, our policies, known as our Community Standards,⁸ outline what is and is not allowed on Meta's services. These policies are developed based on a range of values to help combat abuse. Safety is a core value of our Community Standards, alongside privacy, authenticity, voice, and dignity.⁹

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

⁸ See Meta, *Community Standards*, <https://www.facebook.com/communitystandards>

⁹ Monika Bickert, *Updating the values that inform our community standards*, <https://about.fb.com/news/2019/09/updated-the-values-that-inform-our-community-standards/>

Our policies are 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 people and communities that might otherwise be overlooked or marginalised.

Our Community Standards are regularly updated to keep pace with changes happening online and offline around the world. Every two weeks, members of our Product Policy team run a meeting called the Product Policy Forum to discuss potential changes to our Community Standards, ads policies and major News Feed ranking changes. 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.¹⁰ A change log of changes made to each policy area is available within the Community Standards.¹¹

Enforcement

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

We have built up teams of experts who work in this space. We now have over 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 action the content consistent with our policies. However, increasingly, we have been investing in proactive detection technology to identify and action harmful content before anyone sees it and needs to report 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

¹⁰ See, for example, <https://about.fb.com/news/2018/11/content-standards-forum-minutes/>

¹¹ See, for example Hate Speech: <https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/>

review. We continue to build technologies like RIO,¹² WPIE¹³ and XLM-R¹⁴ that can help us identify harmful content faster, across languages and content type (i.e. text, image, etc.). These efforts and our continued focus on AI research help our technology scale quickly to keep 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.¹⁵ The report is released quarterly and includes metrics such as how much content we are actioning, and what percentage was detected proactively. We now report on 14 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 Community Standards. 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 more than 99 per cent across serious content types such as child exploitation material, terrorist content, violent and graphic content, and fake accounts.

We have led the industry in developing transparency reports about content enforcement, particularly in relation to one of the most significant metrics we provide in our Community Standards Enforcement Report: *prevalence*. Prevalence measures the number of views of violating content, divided by the estimated number of total content views on Facebook or Instagram.¹⁶

¹² 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/>

¹³ 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/>

¹⁴ 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/>

¹⁵ Meta, *Community Standards Enforcement Report* <https://transparency.fb.com/data/community-standards-enforcement/>

¹⁶ Meta, *Prevalence*, *Transparency Centre*, <https://transparency.fb.com/en-gb/policies/improving/prevalence-metric/>

Prevalence is a vital metric because the way content causes harm on the internet is by being seen. Given the nature of the internet, the amount of times content is seen is not evenly distributed. A small amount of content could go viral and get a lot of distribution in a very short span of time, whereas other content could be on the internet for a long time and not be seen by anyone. It is important this distinction is taken into consideration when understanding how we enforce harmful content.

Prevalence is an objective metric that helps policymakers and the community understand how much violating content is actually being seen on our services.

Tools and products

Third, we build technology to help prevent abuse and harmful experiences in the first place, and also design tools to give people more control and help them stay safe. We believe people should have tools to customise their experience on our services - even if content does not violate our policies, people may still find it objectionable or may choose not to see it.

In addition to the long-standing tools of Block, Report, Hide, Unfollow,¹⁷ 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. Our tools aim to discourage harmful behaviour, help users control their experience, and guide users to authoritative information. For example, we provide tools to help users:

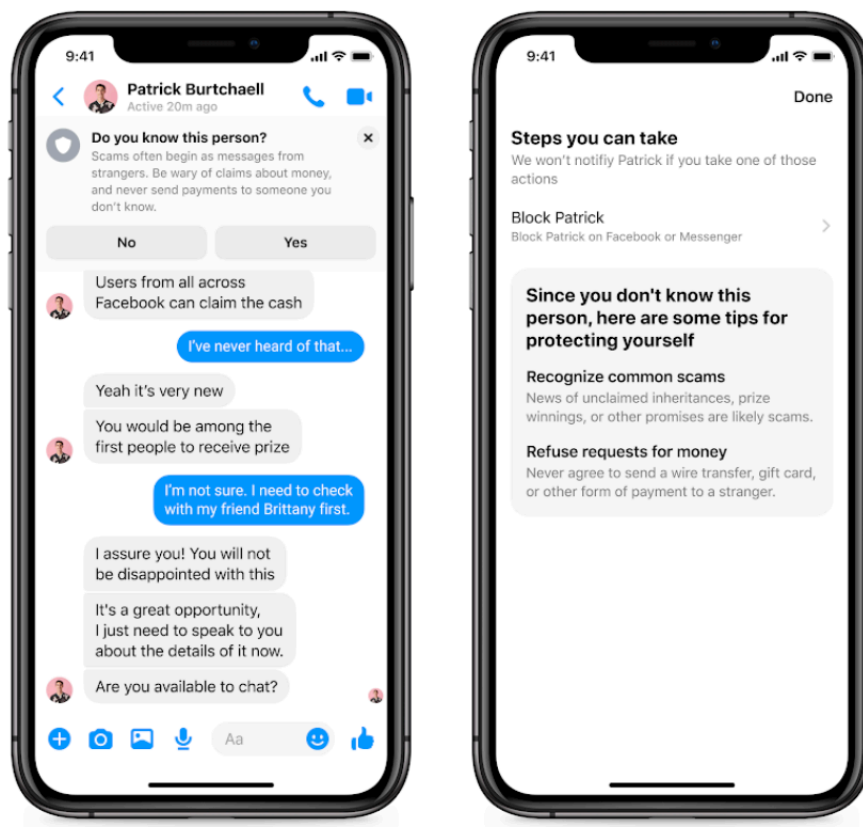
- **Discourage harmful behaviour.** We've introduced warnings and safety notices across our platforms to educate people on who they're engaging with. For example, in Messenger we have introduced safety notices that pop up and provide tips to help people spot suspicious activity or take action to block or ignore someone when something doesn't seem right, shown in Figure 1 below.¹⁸ These notices are designed to discourage inappropriate interactions with children and to limit the potential for grooming to occur via Messenger and Instagram.¹⁹

¹⁷ An overview of these and other tools is available in the Facebook Safety Center:
<https://www.facebook.com/safety/tools>

¹⁸ 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/>

¹⁹ 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/>

Figure 1: Messenger Safety Notice



- **Help users customise and control their experience.** Users can manage the comments they see by ignoring, deleting or restricting unwanted interactions. We also enable users to control who can tag them, or who can send them direct messages.²⁰ A number of other tools to help users customise their experience are outlined in further detail in the ‘Bullying and Harassment’ and ‘Public Figures’ sections below.
- **Guide users to authoritative advice and support.** Throughout our platform, we make resources available at appropriate “just-in time” points. For example, if someone searches for “domestic violence”, they are directed to expert advice and resources. Further, while we don’t allow content that promotes or encourages self-harm and eating disorders, we do allow people to share their own experiences

²⁰ A Davis, Our commitment to keeping people safe, *Meta Newsroom*, 11 February 2020, <https://about.instagram.com/blog/announcements/making-instagram-safer-for-the-youngest-members-of-our-community>

and journeys around self-image and body acceptance. We know that these stories can prompt important conversations and provide community support, but can also be triggering for some. To address this, when someone tries to search for or share self-harm related content, we currently blur potentially triggering images and point people to helpful resources. This includes directing people to dedicated resources, including in Australia, the Butterfly Foundation.²¹

Resources

Fourth, we provide informative resources and learning modules for our users to raise awareness of online safety, and the tools available to help them manage their experience. This includes the Instagram Safety and Wellbeing Hub²² and the Facebook Safety Center.²³ These also include the:

- Bullying Prevention Hub developed in partnership with the Yale Centre for Emotional Intelligence;
- Youth Portal which provides a central place for teens to access education on our tools and products, first person accounts from teens about how they're using technologies, tips on security and reporting, and advice on how to use social media safely;²⁴
- Get Digital Hub, a digital citizenship and wellbeing program which provides schools and families with lesson plans and activities to help build the core competencies and skills young people need to navigate the digital world in safe ways;²⁵
- Suicide Prevention Support Centre that provides resources and guidance on how to access and offer support.²⁶

We have included more detail on dedicated resources we have developed with our partners in the relevant sections below.

²¹ Instagram, how we're supporting people affected by eating disorders and negative body image, *Meta Newsroom*, 22 February 2021, <https://about.fb.com/news/2021/02/supporting-people-affected-by-eating-disorders-and-negative-body-image/>

²² Instagram, *Instagram Community*, <https://about.instagram.com/community>

²³ Meta, *Digital Literacy Library*, <https://www.facebook.com/safety/educators>

²⁴ Meta, *Youth Portal*, https://www.facebook.com/safety/youth?locale=en_GB

²⁵ Meta, *Get Digital Hub*, <https://www.facebook.com/fbgetdigital>

²⁶ Meta, *Suicide Prevention Support Centre*, <https://www.facebook.com/safety/wellbeing/suicideprevention/>

Partnerships

Finally, we have over 400 safety partners across the world, including a number of partnerships in Australia, to ensure our global safety efforts are complemented by on-the-ground expertise.

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

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.²⁷ 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 nine-year partnership that has directly empowered more than 25,000 young Australians to tackle cyberbullying.²⁸

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 the Alannah and Madeline Foundation, as well as many others.

In addition, during the course of the pandemic, we have provided significant support to our safety partners to ensure that our users – especially young people – can connect and communicate safely. Specifically, we have funded and supported the following initiatives:

- **The Alannah and Madeline Foundation.** Delivered a research paper into the impacts of social distancing and isolation on young people, and how technology can alleviate challenges.
- **Cyber Safety Solutions.** Led by Susan McLean, Cyber Safety Solutions delivers online education to students and parents across Australia. We supported continued education and resources for parents in a new online format.
- **Orygen.** We partnered with Orygen, the youth mental health body, on two initiatives – launching #SafeSpace on 2021's World Suicide Prevention Day, and to amplify their #ChatSafe guidelines.²⁹

²⁷ Project Rockit, *Launching: Digital Ambassadors*, <https://www.projectrockit.com.au/digitalambassadors/>

²⁸ R Thomas, 'Young People at the Centre', *Facebook Australia blog*, 8 February 2021
<https://australia.fb.com/post/young-people-at-the-centre/>

²⁹ A Davis, Creating hope through action for suicide prevention and awareness, *Meta Newsroom*, 22 September 2021, <https://about.fb.com/news/2021/09/creating-hope-through-action-for-suicide-prevention-and-awareness/>

#SafeSpace is an interactive digital space that spotlights real stories shared from young people with lived experiences. The stories show how recovery is possible, as well as the importance of normalising conversations around mental health topics.

#ChatSafe provides guidelines to support those who might be responding to suicide-related content, or for those who might want to share their own feelings about suicidal thoughts, feelings or behaviours. With the increase of online learning due to COVID-19, we added Orygen's #chatsafe guidelines to Facebook's Safety Centre. These Guidelines also won the Suicide Prevention Australia's 2021 LIFE Aware for Innovation, acknowledging the integral work Orygen is doing to support safe communication online.

Meta was also one of the first companies to publicly support the Office of the Australian eSafety Commissioner's Safety By Design framework.³⁰

In July 2019, we partnered with a number of organisations - including the eSafety Commissioner's Office - to deliver a Safety by Design Jam. The Safety by Design Jam was a workshop designed specifically for young people, which sought to gather insights and feedback from the people best placed to talk about youth safety online - young people themselves. It was the first of five age-appropriate Design Jams around the world, aiming to bring together policymakers, academics, safety and privacy experts and young people, to share new ideas on how we can build age-appropriate experiences on our platforms to meet the needs of our young community. The process resulted in a publicly available guide with Safety by Design guidelines for developing apps for young people.³¹

In addition to the overview of our approach to safety, we outline below our approach to safety in relation to particular groups within our community in respect of which there has been recent public debate - young people and parents, women's safety and public figures.

³⁰ Safety by Design Youth Jam, *Facebook*, August 2019, <https://www.facebook.com/MetaAustralia/videos/910843179301219/>

³¹ Trust, Transparency and Control Labs, How to design with trust, transparency and control for young people, *TCC Labs*, March 2021, <https://www.ttclabs.net/report/how-to-design-with-trust-transparency-and-control-for-young-people>

Supporting young people and parents

The safety of young people on our services is of paramount importance to us. We want them to have an experience that is both fun and safe. And we want to support their parents to assist them in doing this.

Creating an experience on Facebook and Instagram that's 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. And we have also engaged with parent groups to better understand the resources they need.

We synthesised the outcomes of the 2019 Safety By Design Jam (mentioned above) with additional consultations with the US Federal Trade Commission, UK Information Commissioner's Office, academics, civil rights groups, and industry, into a Youth Design Guide.³² The Guide was co-developed with Trust Transparency and Control Labs and provides advice for product designers and developers. It suggests any products used by young people should follow the principles of (1) designing for different levels of maturity; (2) empowering young people with meaningful transparency and control; and (3) undertaking data education for young people.

We draw from all of the feedback provided, as well as other best practice resources, to design all our products. For example, we have recently launched our internal Youth Knowledge Library, which includes best practices for product teams throughout the company. The Library is modeled on external guidance and frameworks developed by organisations like the United Nations Convention on Rights of the Child, the OECD, and children's rights groups – as well as our own consultation with third party experts, young people, parents and guardians. This resource ensures we have a company-wide understanding of how to apply general principles to specific experiences in our products, and will provide product teams with guidelines on how to design youth products in an age-appropriate way.

³² Trust, Transparency and Control Labs, 'How to design with trust, transparency and control for young people, *TTC Toolkit*, March 2021, <https://www.ttclabs.net/research/how-to-design-with-trust-transparency-and-control-for-young-people>

Bullying and harassment

One of the issues that can be faced by people online, and in particular young people where parents need greater support, is bullying and harassment. Often this may be initiated or may also occur offline, and the online bullying and harassment is simply an extension.

When it comes to bullying and harassment, context and intent matter. Bullying and harassment are often very personal — it shows up in different ways for different people. We therefore continue to update our policies, enforcement, tools and partnerships to ensure our approach to combatting bullying online remains up to date and effective.

We use human review and developed AI systems to identify many types of bullying and harassment across our platforms. However, as mentioned above, because bullying and harassment is highly personal by nature, using technology to proactively detect these behaviours can be more challenging than other types of violations. It can sometimes be difficult for our systems to distinguish between a bullying comment and a light-hearted joke without knowing the people involved or the nuance of the situation. That's why we also rely on people to report this behaviour to us so we can identify and remove it.

Our latest Community Standards Enforcement Report outlines the significant progress we have made in removing bullying and harassment material. In the third quarter of 2021:³³

- We actioned 9.2 million pieces of content on Facebook for violating our policies on bullying and harassment, and of that, 59.4 per cent of bullying and harassment content was removed proactively via artificial intelligence. This is an increase from 54.1 per cent in the previous quarter, and 25.9 per cent one year prior.
- We actioned 7.8 million pieces of bullying and harassment content on Instagram, and of that, 83.2 per cent of it was removed proactively. This is an increase from 71.5 percent in the previous quarter and 54.5 per cent one year prior.

In the third quarter this year, we began reporting on the prevalence of bullying and harassment.³⁴ Prevalence across the platforms was 0.14-0.15 per cent on Facebook and 0.05-0.06 per cent on Instagram. This means bullying and harassment content was seen between 14 and 15 times per every 10,000 views of content on Facebook, and between 5 and 6 times per 10,000 views of content on Instagram. This metric captures only bullying

³³ Meta, *Community Standards Enforcement Report Q3 2021 - Bullying and harassment*, <https://transparency.fb.com/data/community-standards-enforcement/bullying-and-harassment/facebook/>

³⁴ G Rosen, Community Standards Enforcement Report, Third Quarter 2021, *Meta Newsroom*, 9 November 2021, <https://about.fb.com/news/2021/11/community-standards-enforcement-report-q3-2021/>

and harassment where we do not need additional information such as a report from the person experiencing it to determine if it violates our policy.

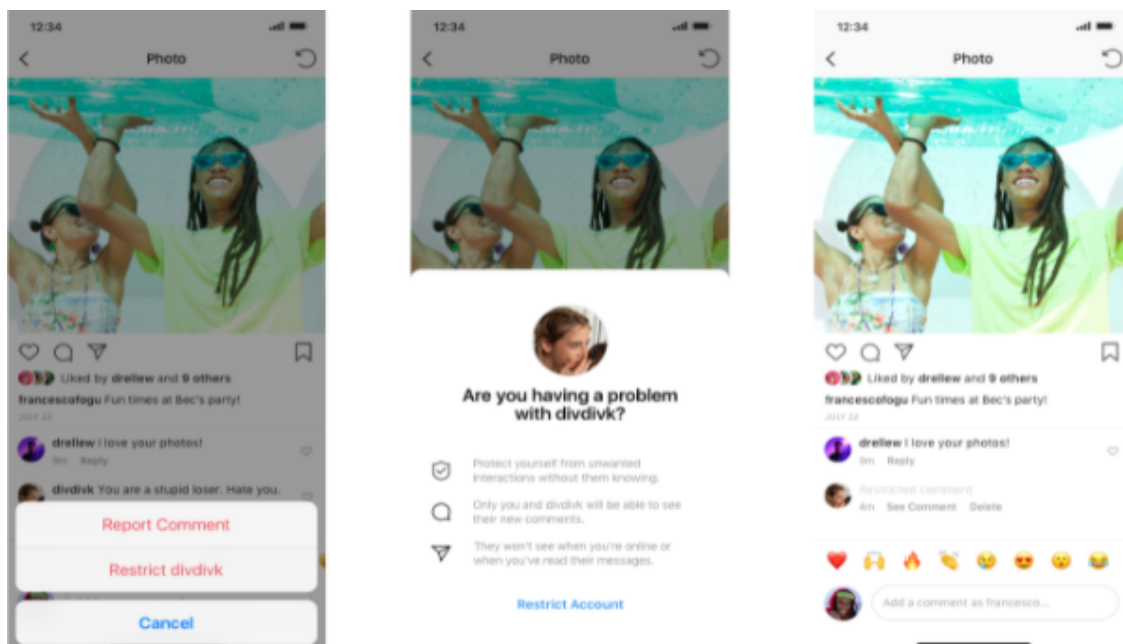
Tools

Even if content does not violate our Community Standards, people may prefer to not see it. They may also want to take steps in order to control their individual experience on our platform.

As mentioned above, as well as our longstanding tools of Block, Report, Hide, Unfollow we've invested in a range of other industry-leading tools including:

- Restrict tool.** We've created a Restrict tool in Instagram³⁵, shown in Figure 2 below, where comments on your posts from a person you have restricted will only be visible to that person. Direct messages will automatically move to a separate Message Requests folder, and you will not receive notifications from a restricted account. You can still view the messages but the restricted account will not be able to see when you've read their direct messages or when you are active on Instagram.

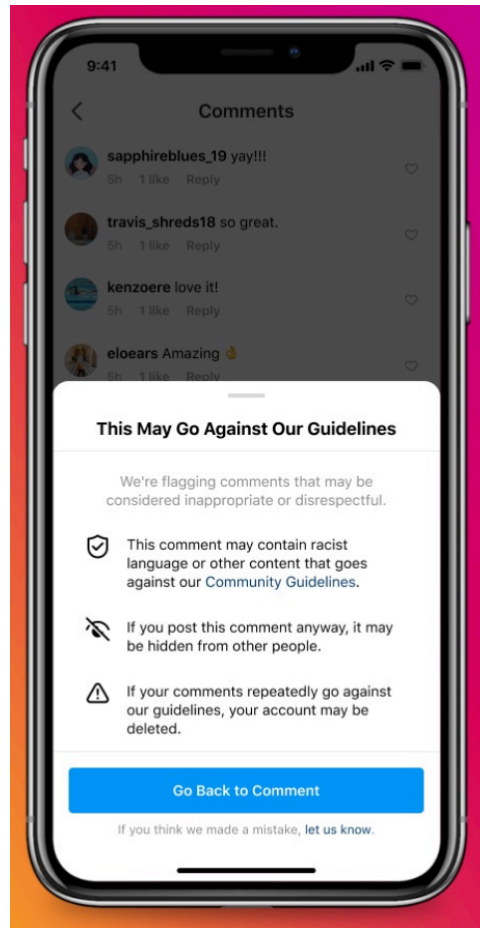
Figure 2: Instagram 'Restrict' tool



³⁵ Instagram, 'Introducing the "Restrict" Feature to Protect Against Bullying', *Instagram Blog*, 2 October 2019, <https://about.instagram.com/blog/announcements/stand-up-against-bullying-with-restrict>.

- **Bullying and harassment warning.** One recent tool we've deployed on both Facebook and Instagram is sending a warning to educate and discourage people from posting or commenting in ways that could be bullying and harassment, shown in Figure 3 below. We've found that after viewing these warnings on Instagram, about 50 per cent of the time the comment was edited or deleted by the user.³⁶

Figure 3: Warnings to discourage bullying or harassment



- **Stopping people tagging or mentioning teens that don't follow them.** We enable users to switch off the ability for people to tag or mention teens who don't follow them, or to include their content in Reels Remixes or Guides by default when they first join Instagram.

³⁶ A Davis, Our approach to addressing bullying and harassment, *Meta Newsroom*, 9 November 2021, <https://about.fb.com/news/2021/11/how-meta-addresses-bullying-harassment/>

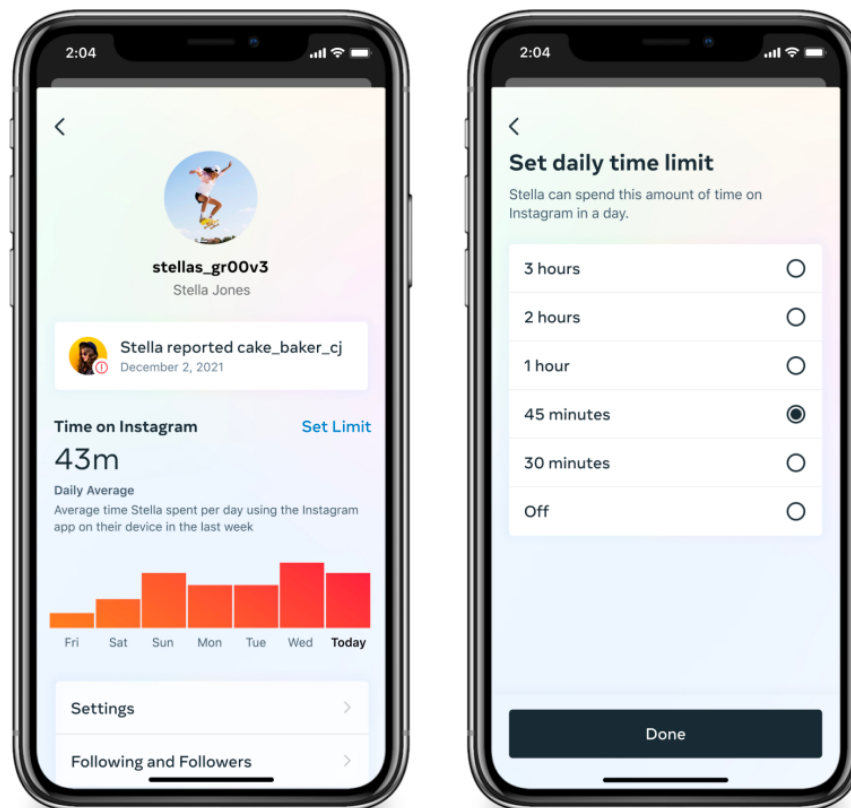
In addition to the responsibility of industry to invest in safety, parents and guardians play a vital role in ensuring the safety of young people online. We want to provide tools and resources for parents and guardians so they can guide and support their teens.

Parental controls are built into our services depending on the nature of the service. For example, in 2020 we launched Messenger Kids in Australia which places parental controls at the heart of the experience so younger users can connect with their friends, while parents can monitor the privacy and security controls. Further detail on Messenger Kids is provided in the 'Ensuring Age-Appropriate Experiences' section below.

In December 2021, we announced we are introducing new tools on Instagram that will provide greater controls for parents. Parents and guardians will soon be able to view how much time their teens spend on Instagram and set time limits, shown in Figure 4 below.³⁷ We'll also give teens a new option to notify their parents if they report someone, giving their parents the opportunity to talk about it with them. This is the first version of these tools and we'll continue to add more options over time.

³⁷ A Mosseri, Raising the standard for protecting teens and supporting parents online, *Meta Newsroom*, 7 December 2021, <https://about.fb.com/news/2021/12/new-teen-safety-tools-on-instagram/>

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

- **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.³⁸ We are also in the process of developing a new educational hub for parents and guardians that will include product tutorials and tips from experts, to help them discuss social media use with their teens.

³⁸ Meta, *Parents Portal*, <https://www.facebook.com/safety/parents>

- **Parents Guide to Instagram.** In Australia, we worked with ReachOut to develop an Instagram Parents' Guide to support parents in understanding Instagram's safety tools. The Guide contains tips for parents to understand Instagram's safety features and how to have 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.³⁹ The Guide was first released in September 2019 and updated in June 2021.⁴⁰

During the 2021 National Child Protection Week, we also held a virtual panel with the Carly Ryan Foundation NSW Police and ReachOut to discuss tools, tips and resources for parents to keep their children safe online.⁴¹

Women's safety

Women come to Facebook and Instagram to run thriving businesses, support each other through Groups and make donations to causes they are passionate about. However, like society, it can also be a place where women experience a disproportionate level of harassment and abuse. It is important that we implement tailored approaches to minimise harm for women, and equip women to manage their online experience. The way in which abuse and harassment manifests online varies country by country, however on the whole, women have a less safe experience than men. One of our key priorities is to ensure safety concerns are addressed, and that women have equal access to all of the economic opportunity, education, and social connections the internet can provide.

We take a comprehensive approach to making our platform a safer place for women including tailored policies and developing cutting-edge technology to help prevent abuse from happening in the first place.

Our approach is informed by consultations with women's safety organisations, industry and experts. Since 2016, we have convened over 200 organisations and experts in

³⁹ Reach Out, *A parents guide to Instagram*, <https://parents.au.reachout.com/landing/parentsguidetoinsta>

⁴⁰ J Machin, 'A Parent's Guide to Instagram', *Facebook Australia blog*, 22 June 2021, <https://australia.fb.com/post/a-parents-guide-to-instagram-in-partnership-with-reach-out/>

⁴¹ M Garlick, Every child, in every community, needs a fair go, *Facebook Australia Blog*, 7 September 2021, <https://australia.fb.com/post/every-child-in-every-community-needs-a-fair-go/>; Meta, National Child Protection Week panel, *Facebook*, 7 September 2021, <https://www.facebook.com/watch/?v=147823834189789>

women's safety roundtables across the world, including Australia. These roundtables inform our ongoing policy development, tools and programs such as the NCII pilot outlined below.⁴²

In 2020, we were one of the first technology companies to appoint a Global Head of Women's Safety, and in 2021 we announced our Global Women's Safety Expert Advisors,⁴³ a group of 12 nonprofit leaders, activists and academic experts to help us develop new policies, products and programs that better support the women who use our apps.. This expert group includes Dr Asher Flynn, an Associate Professor of Criminology at Monash University and the Vice President of the Australian and New Zealand Society of Criminology. Dr Flynn's work focusses on AI-facilitated abuse, deepfakes, gendered violence and image-based sexual abuse.

Policies

In response to feedback we have received from our consultations, we have updated our policies to adjust for the gendered and culturally specific nature that some forms of online harassment and abuse can occur, especially for women. In July 2019, for example, our policy team expanded our bullying and harassment policy to enforce more strictly on cursing that uses female-gendered terms.

Our policies have also been developed to provide more protections for public figures, particularly female public figures, so that they are not subjected to degrading or sexualised attacks. We currently remove attacks on public figures that encompass a wide range of harms. Last year, we announced further changes to this policy to remove unwanted sexualised commentary and repeated content which is sexually harassing.⁴⁴ Because what is "unwanted" can be subjective, we'll rely on additional context from the individual experiencing the abuse to take action. We made these changes because attacks like these can weaponise a public figure's appearance, which is unnecessary and often not related to the work these public figures represent. More details about this update is included below in the discussion about public figures.

⁴² Meta, Making Facebook a safer, more welcoming place for women, *Meta Newsroom*, 29 October 2019, <https://about.fb.com/news/2019/10/inside-feed-womens-safety/>

⁴³ C Southworth, Partnering with experts to promote women's safety, *Meta Newsroom*, 30 June 2021, <https://about.fb.com/news/2021/06/partnering-with-experts-to-promote-womens-safety/>

⁴⁴ A Davis, Advancing our policies on online bullying and harassment, *Meta Newsroom*, 13 October 2021, <https://about.fb.com/news/2021/10/advancing-online-bullying-harassment-policies/>

Tools

Tools such as blocking, reporting and other user-facing tools are only part of the solution for helping women feel safe online. The success of our tools relies on people knowing about them, and understanding and feeling comfortable using them. A victim who's already feeling anxious or threatened may not want to trigger a harasser for fear of retribution. Sometimes, the behaviour isn't visible to the woman it affects: an ex might share non-consensual intimate images in a private group, for example. Or a bully might set up a fake account in a woman's name and operate it without her knowledge, adding members of her community as friends. That's why Meta has not only invested in digital literacy programs and improved safety resources but we have also invested in technology that can find violating content proactively — and in some cases, prevent it from being shared in the first place.

One example of this is our investment in industry-leading initiatives to combat the non-consensual sharing of intimate images (NCII). It has long been our policy on Facebook and Instagram to remove NCII, and in 2017 we began a pilot in 9 countries - including in Australia with the Office of the eSafety Commissioner - to help victims proactively stop the proliferation of their intimate images.⁴⁵

Following the success of this pilot, we recently launched the expansion of the program globally, known as StopNCII.org. StopNCII.org will operate in partnership with more than 50 non-governmental organisations around the world, including the Office of the eSafety Commissioner.

This is the first global initiative of its kind to safely and securely help people who are concerned their intimate images (photos or videos of a person which feature nudity or are sexual in nature) may be shared without their consent.⁴⁶

When someone is concerned their intimate images have been posted or might be posted to online platforms like Facebook or Instagram, they can create a case through StopNCII.org. When they select their image, the tool uses hash-generating technology to assign a unique hash value (a numerical code) to the image, creating a secure digital fingerprint. The original image never leaves the person's device. Only hashes, not the images themselves, are shared with StopNCII.org. Tech companies participating in

⁴⁵ Meta, *Non-consensually shared intimate images pilot*, <https://www.facebook.com/safety/notwithoutmyconsent/pilot/how-it-works>

⁴⁶ A Davis, Strengthening our efforts against the spread of non-consensual intimate images, *Meta Newsroom*, 2 December 2021, <https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images/>

StopNCII.org receive the hash and can use that hash to detect if someone has shared the images or is trying to share those images on their platforms. Creating a case through StopNCII.org can actively stop the proliferation of NCII.

We've developed this platform with privacy and security at every step thanks to extensive input from victims, survivors, experts, advocates and other tech partners. By allowing potential victims to access the hashing technology directly we are giving them more privacy and control of their images.

Resources

We work with third party experts to develop resources specifically designed to promote women's safety, these include:

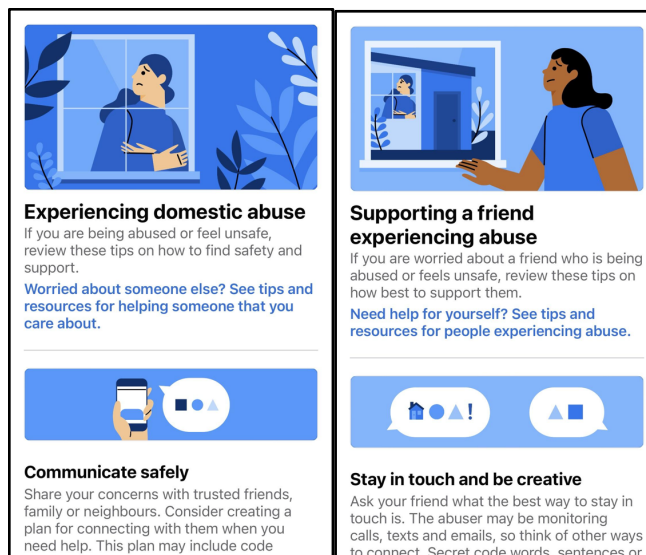
- Not Without My Consent, which provides information about our recently announced global program, StopNCII.org,⁴⁷ that helps victims proactively stop the proliferation of their intimate images.
- The Stop Sextortion Hub, which we have developed with global NGO Thorn, with resources for teens, caregivers and educators seeking support and information related to sextortion.
- A dedicated safety page for women on our Safety Centre Hub.⁴⁸

Amidst growing concerns about increased domestic violence during COVID-19, Meta has focused on ensuring people are able to easily connect with trained representatives from Australian helplines. We compiled a list of resources around the world in partnership with UN Women, the U.S. National Network to End Domestic Violence and the Global Network of Women's Shelters. These are all included on the COVID-19 Information Centre, under modules called "Supporting a Friend Experiencing Abuse Tips" and "Experiencing Domestic Abuse", shown in Figure 5 below. We have also sent untargeted prompts to 1 million users directing people to these resources.

⁴⁷ A Davis, Strengthening our efforts against the spread of non-consensual intimate images, *Meta Newsroom*, 2 December 2021, <https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images/>

⁴⁸ Meta, *Women's Safety Hub*, <https://www.facebook.com/safety/womenssafety>

Figure 5: COVID-19 Information Hub Domestic Abuse Module



Partnerships

Finally, we have launched a number of partnerships in Australia to ensure our global safety efforts are complemented by on-the-ground expertise and knowledge. These include:

- As mentioned in the ‘Safety’ section above, we work closely with the Alannah and Madeline Foundation. Specifically in 2019, we worked with the Alannah and Madeline Foundation and the Stars Foundation to create a new program, Safe Sistas,⁴⁹ which supports the online safety of young Indigenous women to respond to the issue of non-consensually shared intimate images in a culturally relevant and safe way. The program reached 857 young girls in Years 7 - 12 in remote and regional communities across NT, QLD and VIC.

A program evaluation conducted by Macquarie University’s Department of Indigenous Studies⁵⁰ found the program found:

- All students reported having an increased confidence in their ability to protect their privacy and online identity when it comes to sharing images online as a result of participating in the Safe Sistas workshop

⁴⁹ Alannah & Madeline Foundation, *Helping Sistas be safer*, <https://www.amf.org.au/news-events/latest-news/helping-sistas-be-safer/>

⁵⁰ Department of Indigenous Studies, Macquarie University, *Safe Sistas Evaluation Report* <https://researchers.mq.edu.au/en/publications/safe-sistas-evaluation-report>

- 73% of students reported that the Safe Sistas workshop has helped them be more likely to think before they act when it comes to sharing images and content online.
- During COVID-19, we also worked with WESNET, 1800 RESPECT and Our Watch to promote their messaging campaigns around family violence, so they can reach vulnerable Australians who may need their help.
- In September 2021, Meta also hosted an online discussion with Minister for Communications Paul Fletcher, Dr Asher Flynn and Cindy Southworth, Global Head of Women's Safety Policy at Meta, together with Mamamia, to raise the profile of work being done by Government, industry and community to support women's safe experience online.⁵¹ The event reached more than 32,000 people.

Public figures

It is important that all users feel safe and protected on our platforms, including those with a public life who use our services to engage and connect with their communities. As noted in the 'Women's safety' section above, whilst our Bullying and Harassment policy distinguishes between public figures and private individuals, we do remove content directed at public figures that contains hate speech, or contains gendered or sexualised attacks, or other severe attacks. We also recognise the volume of engagement and comments that public figures can receive, and have continued to build out tools to enable them to manage this and their exposure to harmful comments.

Policies

We regularly update our policies to reflect society's expectations and feedback from experts and stakeholders, and we recently updated our policies to increase enforcement against harmful content for public figures.⁵² These updates came after years of consultation with free speech advocates, human rights experts, women's safety groups, cartoonists and satirists, female politicians and journalists, representatives of the LGBTQIA+ community, content creators and other types of public figures.

⁵¹ Meta, Women's Safety Panel, *Facebook*, 22 September 2021, https://www.facebook.com/watch/?extid=NS-UNK-UNK-UNK-IO5_GKOT-GK1C&v=266923491951411

⁵² A Davis, Our approach to addressing bullying and harassment, *Meta Newsroom*, 9 November 2021, <https://about.fb.com/news/2021/11/how-meta-addresses-bullying-harassment/>

First, as noted, we've expanded our protections for public figures to include the removal of severe or unwanted sexualising attacks.

Second, recognising that not everyone in the public eye chooses to become a public figure but can still be the subject of bullying and harassment, we've increased protections for involuntary public figures, like human rights defenders and journalists. For example, content that attacks the appearance of a woman journalist would violate our policies and be enforced against.

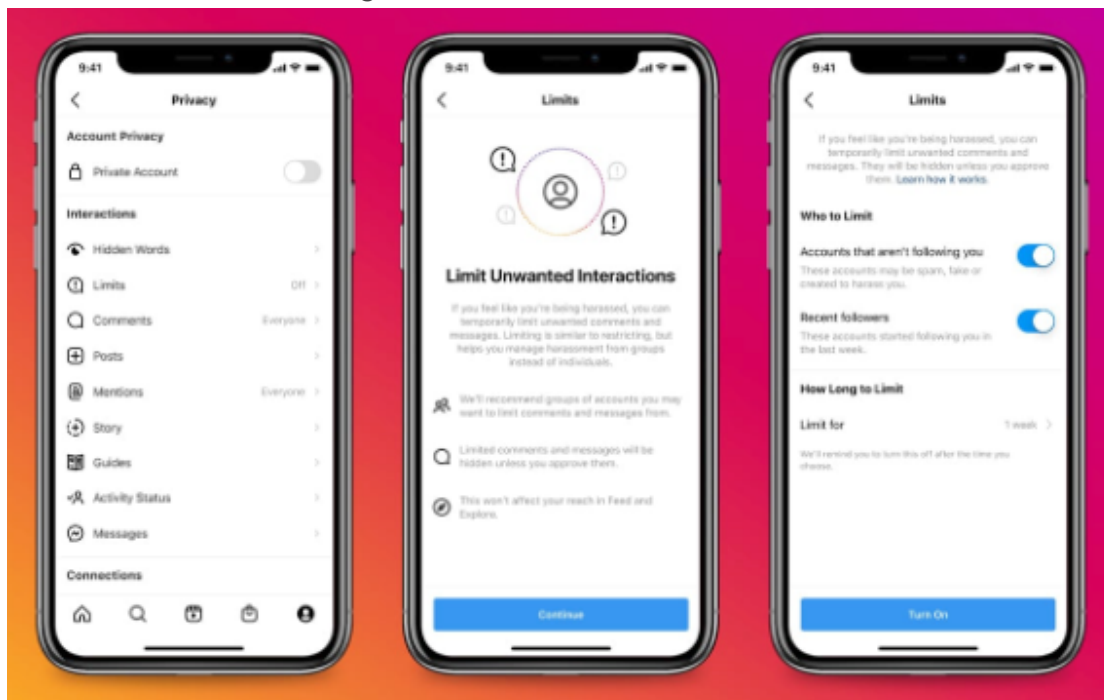
Tools

In consultation with experts and public figures themselves, we have introduced a number of specific tools that help users reduce unwanted interactions online, including:

- **Limits.** The Limits tool on Instagram, shown in Figure 6 below, allows users to automatically hide comments and Direct Messaging requests from people who don't follow them, or who only recently followed them, to help to manage an unexpected rush of unwanted contact.⁵³ In Australia we developed and launched this tool in partnership with the Australian Football League (AFL), to help protect their players from racist abuse.

⁵³ Instagram 'Introducing New Ways to Protect our Community from Abuse', *Instagram Blog*, 10 August 2021, <https://about.instagram.com/blog/announcements/introducing-new-ways-to-protect-our-community-from-abuse>

Figure 6: 'Limits' tool on Instagram



- **Restrict commenting audience.** We have introduced tools to give users control over who comments on their posts on Facebook News Feed. Users can control their commenting audience for a post by choosing from a menu of options. By adjusting the commenting audience, users can control how they want to invite conversation onto their public posts, and limit potentially unwanted interactions.⁵⁴
- **Comment Controls.** The Comment Controls feature on Instagram allows users to automatically hide comments based on a list of words, phrases, numbers or emojis that they can manually add to based on their experiences or preferences.⁵⁵ If people comment using those words or emojis, the user will not be notified and they will not be published on the post for anyone to see. We know from research that, while people don't want to be exposed to negative comments, they want more transparency into the types of comments that are hidden. You can tap "View Hidden Comments" to see the comments. Comments that violate our Community Guidelines will continue to be automatically removed.

⁵⁴ R Sethuraman, More control and context in News Feed, *Meta Newsroom*, 21 March 2021, <https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/>

⁵⁵ Instagram, Kicking Off National Bullying Prevention Month With New Anti-Bullying Features, *Instagram Blog*, 6 October 2020, https://about.instagram.com/en_US/blog/announcements/national-bullying-prevention-month.

- **Hidden Words.** We have recently introduced a tool which will automatically filter direct message (DM) requests containing offensive words, phrases and emojis.⁵⁶ This tool focuses on DM requests, because this is where people usually receive abusive messages - unlike the regular DM inbox - where you receive messages from friends. We have worked with leading anti-discrimination and anti bullying organisations to develop a predefined list of offensive terms that will be filtered from DM requests. Users also have the option to create their own custom lists of words, phrases or emojis that they don't want to see in their DM requests, because we know that different words can be hurtful to different people.
- **New blocking features.** To protect users from unwanted contact, last year we launched new blocking features so that whenever you decide to block someone on Instagram, you'll also have the option to block new accounts that person may create.⁵⁷ This is designed to help make sure users don't hear from people they've blocked, even when they create a new account. This is in addition to our harassment policies, which already prohibit people from repeatedly contacting someone who doesn't want to hear from them.

Partnerships

We continue to engage with policymakers, public figures and stakeholders in Australia on online safety for public figures. Most recently, in September 2021, Meta participated in a panel event for the Parliamentary Friends of Making Social Media Safer alongside the eSafety Commissioner, in order to help raise awareness among Australian Parliamentarians about tools that are available to help keep them safe online.

Recognising that many athletes provide strong role models for young Australians and within our community more broadly, we have also been working closely with sporting organisations such as the AFL. Most recently, in December 2021, Meta worked with the AFL to deliver a specialised education workshop for AFL men's and women's players to understand the tools and resources available to them and provide an additional layer of support through peak season moments. This workshop included participation by the eSafety Commissioner's Office.

⁵⁶ Instagram, Introducing new tools to protect our community from abuse, *Instagram Blog*, 21 April 2021, <https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse>

⁵⁷ Ibid.

Ensuring age-appropriate experiences online

We recognise the role that proportionate and risk-based age assurance regulation (in addition to other safety and privacy safeguards) can play in helping to ensure that young people have an age-appropriate experience 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. Below, we outlined the suite of measures we use to understand a user's 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.

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 multiple signals. We look at things like people wishing a user happy birthday and the age written in those posts: for example, "Happy 21st Birthday!". We also look at the age users have shared across apps: for example, if a user has shared their birthday on Facebook, we'll use the same for linked accounts on Instagram.

We're focused on using existing data to inform our artificial intelligence technology. Where we do feel we need more information, we're developing a menu of options for someone to prove their age. This is a work in progress.

We're also in discussions with the wider technology industry on how we can work together to share information in privacy-preserving ways that helps apps establish whether people are over a specific age. One area we believe has real promise is working with operating system (OS) providers, internet browsers and other providers so they can share information to help apps establish whether someone is of an appropriate age.

This has the dual benefit of helping developers keep underage people off their apps while removing the need to go through differing and potentially cumbersome age verification processes across multiple apps and services. While it's ultimately up to individual apps and websites to enforce their age policies and comply with their legal obligations, collaboration with OS providers, internet browsers and others would be a helpful addition to those efforts.

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.

Age-appropriate controls

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

- **Encouraging existing teen accounts to be private.** For young people who already have a public account on Instagram, we will show them a notification highlighting the benefits of a private account and explaining how to change their privacy settings. We'll still give young people the choice to switch to a private account or keep their current account public if they wish.
- **Limiting advertisers' ability to reach young people.** We now only allow advertisers to target ads to people under 18 (or older in certain countries) based on their age, gender and location. This means that previously available targeting options, like those based on interests or on their activity on other apps and websites, will no longer be available to advertisers.

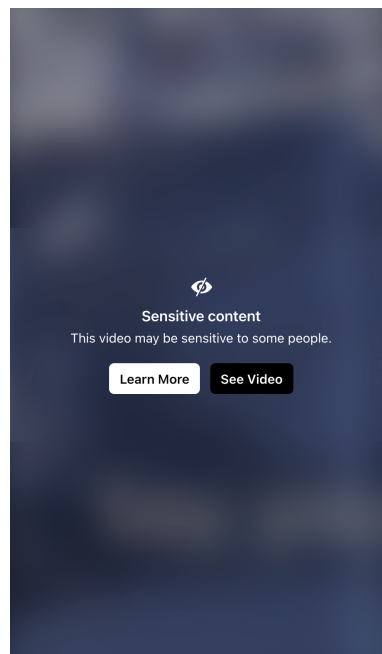
This is in addition to age-gating controls made available for those advertisers who publish age-sensitive ads or content (such as related to gambling).

We already give people ways to tell us that they would rather not see ads through controls within our ad settings. But we've heard from youth advocates that young people may not be well equipped to make these decisions. For this reason, we are taking a more precautionary approach in how advertisers can reach young people with ads.

- **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



- **Restricting adults from privately messaging young people.** As explained above, we send safety notices to users in Messenger, and subsequently Instagram, if we believe an adult could be pursuing a potentially inappropriate private interaction with a teen, see Figure 1 above.⁵⁸
- **Making it more difficult for adults to find and follow teens.** We’ve developed new technology that will allow 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 won’t show young people’s accounts to these adults who exhibit “potentially suspicious behaviour”. If they find young people’s accounts by searching for their usernames, they won’t be able to follow them. They also won’t be able to see comments from young people on other people’s posts, nor will they be able to leave comments on young people’s posts. We’ll continue to look for additional places where we can apply this technology.

⁵⁸ 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/>

These changes are being rolled out in Australia and a small number of other countries initially, and will expand to include other countries soon.

Under 13s

As mentioned above, Facebook and Instagram are designed for users aged 13 and above.

We allow anyone to report suspected underage user accounts on Instagram and Facebook. Our content reviewers are also 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.

Between July and September 2021, Meta removed more than 2.6 million accounts on Facebook and 850,000 accounts on Instagram globally because they were unable to meet our minimum age requirement.⁵⁹

In order to reduce the incentive for users to misrepresent their age, we are also working on providing products and experiences designed specifically for users under 13, managed by parents and guardians.

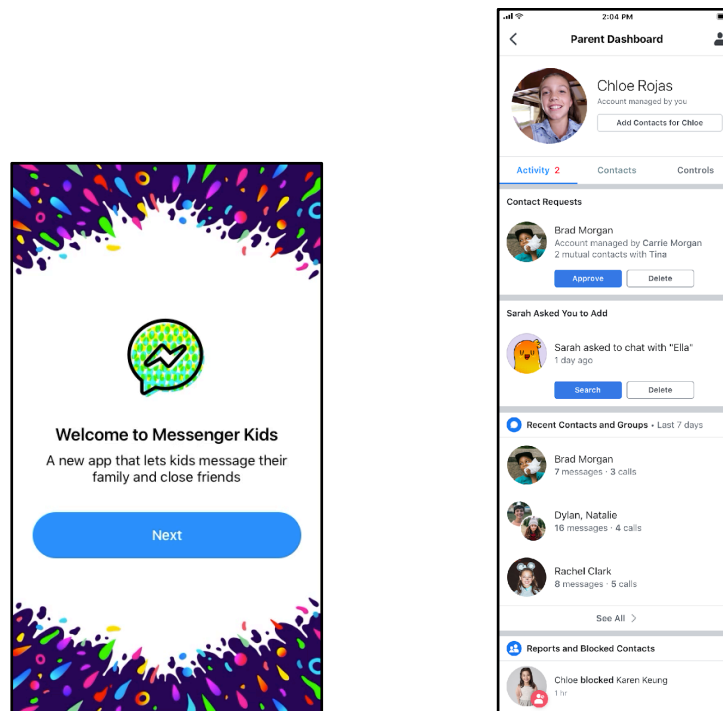
As referenced above, in 2020, in response to the COVID-19 pandemic, we accelerated the launch in Australia of a product called Messenger Kids, shown in Figure 8 below. This is a new messaging product for users who are not yet 13, and provides them with much greater privacy and security controls than regular Messenger. It's a fun way for younger users to connect with their friends, especially while in lockdown or isolation during the pandemic.

Parental control is at the heart of Messenger Kids. Parents manage who their child interacts with and can monitor their child's activity in the app through the Parent Dashboard, where they can also download their child's information at any time.

The design of Messenger Kids, and the control measures, have been developed after extensive consultation with a team of experts in online safety, child development and media, as well as parents.

⁵⁹ A Mosseri, Hearing Before the United States Senate Committee on Commerce, Science, and Transportation Subcommittee on Consumer Protection, Product Safety, and Data Security, 8 December 2021, <https://www.commerce.senate.gov/services/files/3FC55DF6-102F-4571-B6B4-01D2D2C6F0D0>

Figure 8: Screenshots from Messenger Kids, including the Parent Dashboard (right)



As mentioned in the ‘Supporting Young People and Parents’ section above, we’re also developing new ways for parents to supervise their child’s use of Instagram. These new tools will allow parents to oversee their children’s accounts and meaningfully shape their teen’s experience.⁶⁰

⁶⁰ A Mosseri, Raising the standard for protecting teens and supporting parents online, *Meta Newsroom*, 7 December 2021, <https://about.fb.com/news/2021/12/new-teen-safety-tools-on-instagram/>

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 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 your relationships offline, not detracting from them. After all, that’s what apps such as Facebook and Instagram have always been about. This is important as we know that a person’s health and happiness relies heavily on the strength of their relationships.

There is strong evidence that demonstrates the value social media and online interaction can have on a person’s wellbeing. Social media enables people to connect with their family and friends, learn about their passions and interests, and stay up to date with the latest news and trends. It also enables users to find community, connection and supportive spaces. This has been of particular importance over the past two years, as people often have been unable to meet face to face.

However, we 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.

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.⁶¹

Moira Burke, Meta's Data Scientist and Wellbeing Researcher, has undertaken a number of studies on the intersection of wellbeing and social technology.⁶² 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.⁶³

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.

In addition to the research outlined above, we invest in local research in Australia to better understand the experiences of local community groups on our services. Some recent examples of work we have delivered or commissioned include:

- Dr Benjamin Hanckel and Dr Shiva Chandra's recent research with Western Sydney University in Australia shows that while social media can lack diverse content or provide a platform for hate speech, they also provide an environment

⁶¹ 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>

⁶² M Burke, *Latest Publications*, <https://research.facebook.com/people/burke-moira/>

⁶³ Meta, Hard questions: Is spending time on social media bad for us? *Meta Newsroom*, 15 December 2017, <https://about.fb.com/news/2017/12/hard-questions-is-spending-time-on-social-media-bad-for-us/>

for minority groups to find community, connection and supportive spaces.⁶⁴ It finds that social media platforms can empower young people to create online experiences where they feel comfortable and safe; and

- Tristan Kennedy's recent research has found that while Indigenous communities disproportionately experience harmful content online, social media also provides the opportunity for indigenous peoples to express themselves, empowering communities to share stories and speak their truth.⁶⁵

Meta also invests in ongoing research dedicated to mental health and wellbeing, seeking the views of global experts. In 2021, we ran a global funding round for research on safety and community health. This program attached more than 200 proposals from 172 countries, and we are proud that we will support two Australian researchers through this initiative,⁶⁶ including:

- Teddy Cook and Denton Callander from the University of New South Wales - Enhancing trans people's experiences of gender affirmation on Instagram
- Jorge Goncalves, Louise La Sala, Senuri Wijenayake, Simon D'Alfonso from University of Melbourne - Mitigating cyberbullying experiences of younger users on Instagram.

Partnerships

We have a range of partnerships to assist with education and awareness of wellbeing, for both parents and young people:

- We have hosted several **Global Safety and Wellbeing Summits**.⁶⁷ These Summits are joined by over 100 organisations from 40 countries to discuss a wide range of issues including suicide prevention, raising children in the digital era and protecting the most vulnerable people online.

⁶⁴ Dr B Hanckel, Dr S Chandra, 'Social media insights from sexuality and gender diverse young people during COVID-19', *Western Sydney University*, May 2021, https://www.westernsydney.edu.au/_data/assets/pdf_file/0006/1837977/Social_Media_and_LGBTQIA_Youth_Report.pdf

⁶⁵ Dr T Kennedy, 'Indigenous peoples' experiences of harmful content on social media', *Macquarie University*, https://research-management.mq.edu.au/ws/portalfiles/portal/135775224/MQU_HarmfulContentonSocialMedia_report_201202.pdf

⁶⁶ Meta, Announcing the recipients of Instagram research awards on safety and community health, *Meta Research*, 2 December 2021, <https://research.facebook.com/blog/2021/12/announcing-the-recipients-of-instagram-research-awards-on-safety-and-community-health/>

⁶⁷ A Davis, 2019 Global safety and wellbeing summit, *Meta Newsroom*, 16 May 2019, <https://about.fb.com/news/2019/05/2019-global-safety-well-being-summit/>

- **eSafety conferences.** Meta's wellbeing researchers have participated in the annual online safety conferences, jointly organised by the Office of the eSafety Commissioner and Netsafe in 2018 and 2019.
- We supported the **National Mental Health Commission's** to develop the #ChatStarter movement.⁶⁸ To mark the launch of Mental Health week in 2021, the National Mental Health Commission worked with Australia's national mental health organisations - batyr, Beyond Blue, Butterfly Foundation, headspace, Kids Helpline, Orygen, and ReachOut - to create #ChatStarter. The campaign connects, engages, and promotes the benefits of supportive conversations with young people and children who are going through a difficult time. It also encourages young Australians and parents to create their own content on social media with instructions on how they start chats safely with others.
- We have partnered with **headspace**, Australia's national mental health foundation, to host a new online safety and education series for parents in 2021 and 2022. We're working with headspace to run a three-part series on Facebook Live, equipping parents with practical tools to support young people's mental health.
- To support those in LGBTQIA+ communities create a safe and positive online experience, we worked in partnership with **ACON, Twenty10, Black Rainbow, Minus18 and Trans Pride Australia** to create the 'Safe & Strong' guide. This guide provides user-friendly safety tips for Instagram users.⁶⁹
- We developed the #TheWholeMe campaign with the **Butterfly Foundation**. The campaign offers guides for parents and teens to promote positive body image and ensure the wellbeing of young people across our platforms.⁷⁰

Tools and resources

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:

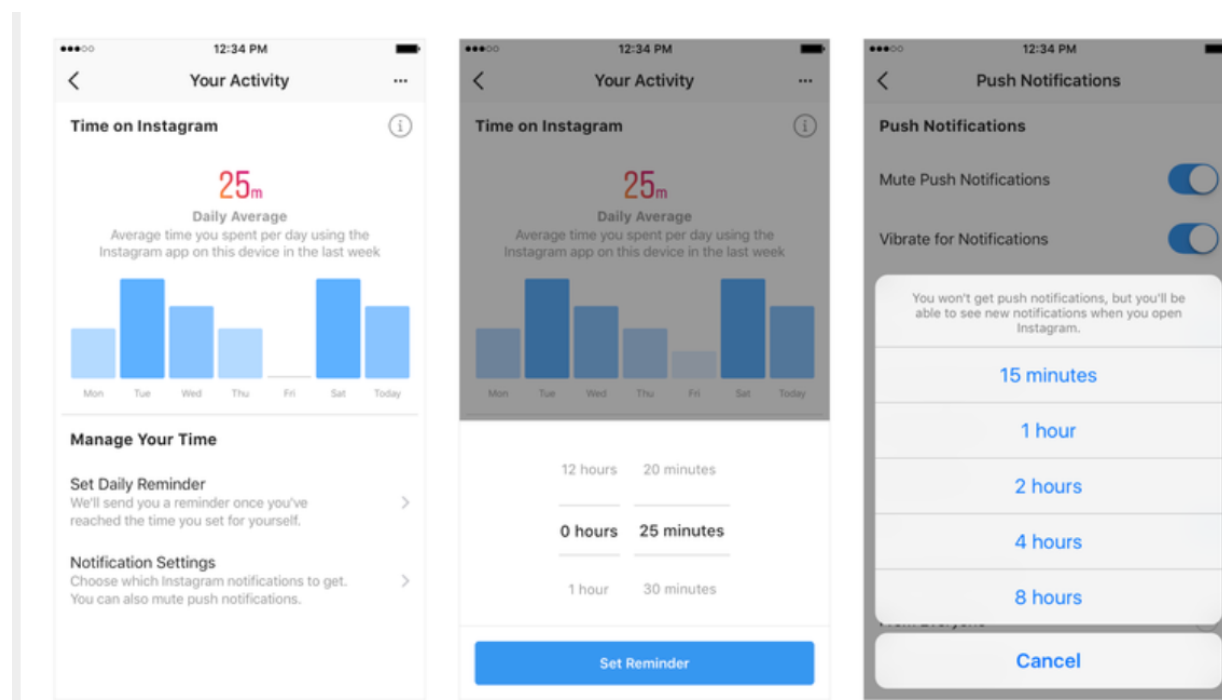
⁶⁸ Christine Morgan, Start a chat: Start a meaningful connection, *Facebook Australia blog*, 11 August 2021, <https://australia.fb.com/post/start-a-chat-start-a-meaningful-connection/>

⁶⁹ ACON, *Instagram Safe and Strong Guide*, <https://www.acon.org.au/wp-content/uploads/2020/02/ACON-Facebook-Instagram-LGBTQ-Guide.pdf>

⁷⁰ The Butterfly Foundation, *The Whole Me Campaign*, <https://butterfly.org.au/get-involved/campaigns/the-whole-me/>

- Improving News Feed quality.** As mentioned above, we've made several changes to News Feed to provide more opportunities for meaningful interactions, and reduce passive consumption of low-quality content.⁷¹ 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 recently 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.⁷²

Figure 9: Activity Dashboard



⁷¹ M Zuckerberg, Meaningful social interaction post, *Facebook*, 2 November 2017, <https://www.facebook.com/zuck/posts/10104146268321841>

⁷² Meta, New tools to manage your time on Facebook and Instagram, *Meta Newsroom*, 1 August 2019, <https://about.fb.com/news/2018/08/manage-your-time/>

- **Suicide prevention tools.** We 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.⁷³

- **Hide Likes on Facebook and Instagram.** We tested hiding like counts to see if it might depressurise people's experience on Instagram.⁷⁴ 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 in their feed. They also have the option to hide like counts on your own posts, so others can't see how many likes your posts get. This way, users who choose to turn off likes can focus on the photos and videos being shared, instead of how many likes posts get.
- **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.⁷⁵ 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. We have launched this feature in the US, UK, Ireland, Canada, Australia, and New Zealand already, and will continue to roll it out to other countries early this

⁷³ G Rosen, Getting our community help in real time, *Meta Newsroom*, 27 November 2017, <https://about.fb.com/news/2017/11/getting-our-community-help-in-real-time/>

⁷⁴ Meta, Giving people more control on Instagram and Facebook, *Meta Newsroom*, 26 May 2021, <https://about.fb.com/news/2021/05/giving-people-more-control/>

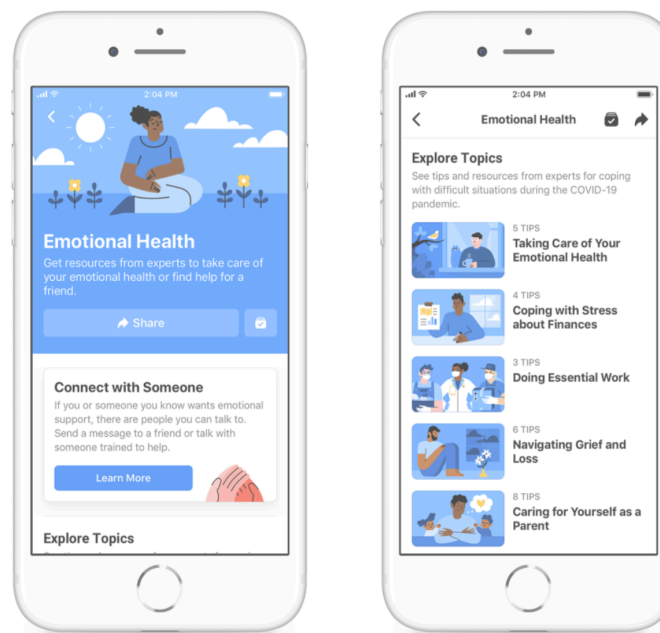
⁷⁵ A Mosseri, Raising the standard for protecting teens and supporting parents online, *Meta Newsroom*, 7 December 2021, <https://about.fb.com/news/2021/12/new-teen-safety-tools-on-instagram/>

year.

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 and Wellbeing Centre⁷⁶ and an Emotional Health Resource Centre⁷⁷ that provide resources on online wellbeing, and transparency around our policies and engagement with experts on the issue. The COVID-19 Information Centre⁷⁸ was also updated to include education modules to maintain your emotional health during the pandemic, see Figure 10 below.

Figure 10: COVID-19 Information Centre Mental Health and Wellbeing Modules



⁷⁶ Meta, *Safety and Wellbeing Centre*, <https://www.facebook.com/safety/wellbeing>

⁷⁷ Meta, Connecting people to mental health resources around the world, *Meta Newsroom*, 5 October 2020, <https://about.fb.com/news/2020/10/connecting-people-to-mental-health-resources/>

⁷⁸ Meta, *COVID-19 Information Centre*, https://www.facebook.com/coronavirus_info

Hate speech

One type of harmful content that we take action on is hate speech. Some claim that the technology industry has an incentive to amplify hate speech online. We address this below, where we provide more detail about the role of algorithms and polarisation. However, first, we wanted to share more detail about our policies, enforcement and partnerships specifically to combat online hate.

Our policies prohibit hate speech on our platforms, and we invest significantly to combat it. Hate speech creates an environment of intimidation and exclusion, may promote offline violence, and can inhibit people from using their voice and feeling safe to connect freely.

Combatting hate speech is a continuous responsibility for governments, working in partnership with experts, industry and the broader community. We take responsibility for detecting and removing hateful communications and groups from our services.

We have outlined below our continued efforts to reduce hate speech on our services through our policies, enforcement efforts, partnerships and research.

Policies

We define hate speech as a direct attack against people on the basis of what we call protected characteristics. We have currently listed the following as protected characteristics:

- race
- ethnicity
- national origin
- disability
- religious affiliation
- caste
- sexual orientation
- sex
- gender identity
- 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. This goes well beyond what is required in Australian legislation.

We have made a number of changes over the last 18 months to expand our hate speech policies in our Community Standards. These include:

- the development of a new hateful stereotypes policy, which will in the first instance prohibit content depicting blackface and stereotypes that Jewish people run the world.⁷⁹ We continue to consult on possible expansions to this policy to capture other hateful stereotypes.
- expansions in our ads policies to better protect immigrants, migrants, refugees and asylum seekers from hateful claims⁸⁰
- expansions in our ads policies to prohibit claims that a group is a threat to the safety, health or survival of others on the basis of that group's race, ethnicity, national origin, religious affiliation, sexual orientation, gender, gender identity, serious disease or disability.⁸¹
- announcing that we will amend our policy to remove any claims that deny or distort the Holocaust, on the basis of expert consultation and research.⁸²

Enforcement

Enforcing our policies against hate speech has traditionally been the most challenging content for artificial intelligence to detect, because it is dependent on nuance, history, language, religion and changing cultural norms. According to the latest Community Standards Enforcement Report, in the period July to September 2021, we took action against 22.3 million pieces of content for hate speech, of which 96.5 per cent of hate speech content was detected proactively.

Our investment in artificial intelligence is evident from the increasing percentage of hate speech content we have been detecting proactively. Hate speech is traditionally one of the most challenging types of content to proactively detect because it is so context-dependent. At the end of 2017, less than 25 per cent of hate speech content we removed was detected proactively. This figure has progressively increased over that time: by end 2018, 40 per cent was proactively detected; by end 2019, 80 per cent was proactively detected (see Figure 1 below).

⁷⁹ G Rosen, 'Community Standards Enforcement Report August 2020', *Meta Newsroom*, 11 August 2020, <https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/>.

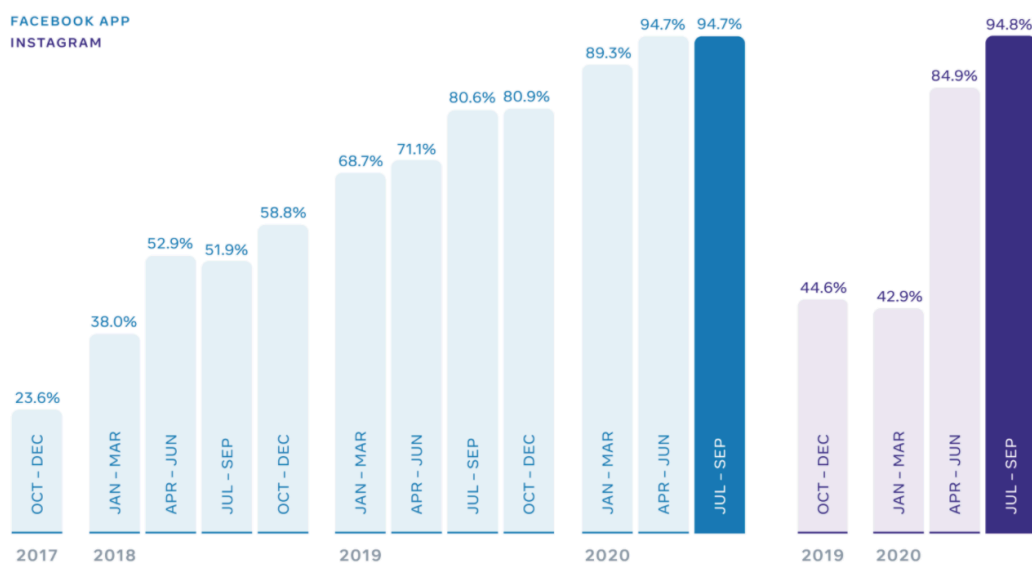
⁸⁰ Meta, 'Meeting the unique challenges of the 2020 elections', *Meta Newsroom*, 26 June 2020, <https://about.fb.com/news/2020/06/meeting-unique-elections-challenges/>

⁸¹ Ibid.

⁸² M Bickert, 'Removing Holocaust denial content', *Meta Newsroom*, 12 October 2020, <https://about.fb.com/news/2020/10/removing-holocaust-denial-content//>

This improvement in our detection ability was accompanied by a stark increase in the total volume of hate speech content we have removed, shown in Figure 11 below. At the end of 2018, we removed 3.4 million pieces of content; at the end of 2019, we removed 5.5 million; and at the end of 2020, we removed 26.9 million. The quantity of hate speech content we have removed has been relatively constant since then.

Figure 11: Hate speech removals on Facebook and Instagram, by percentage of how they were detected



As mentioned above, we also measure how prevalent violating content is on our services. At the end of 2020 0.7 to 0.8 per cent of views on Facebook contained hate speech. This means, for every 10,000 views of content on Facebook, 10 or 11 contained hate speech.⁸³ Now, as reported at the end of 2021, less than 0.03 per cent of views on Facebook contained hate speech.⁸⁴

Our enforcement approach has been scrutinised externally. For example, a 2021 European Commission report found that Facebook assessed 95.7 per cent and Instagram assessed 91.8 per cent of hate speech notifications in less than 24 hours, compared to

⁸³ A Kantor, Measuring our progress combatting hate speech, *Meta Newsroom*, 19 November 2020, <https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/>

⁸⁴ Meta, *Community Standards Enforcement Report - hate speech* <https://transparency.fb.com/data/community-standards-enforcement/>

81.5 per cent for YouTube and 76.6 per cent for Twitter.⁸⁵ The European Commission also stated that “only Facebook informs users systematically; all the other platforms have to make improvements.”

Partnerships

While we have made significant progress as a company in combatting online hate, our work is enriched by partnerships with other companies, civil society organisations, experts, and governments.

For example, over the past twelve months, we have continued building partnerships with Australia-based organisations. These include ongoing engagement with representatives from the Australian Jewish and Muslim communities to seek feedback on what they are seeing in relation to anti-Semitism and Islamophobia.

We also established an Australia-specific Combatting Online Hate Advisory Group in October 2020. The Advisory Group contains representatives of marginalised communities, and experts in different forms of online hate such as white supremacy. The Advisory Group meets quarterly to discuss how industry and civil society can work together to combat online hate in Australia.

⁸⁵ G Rosen, ‘New EU report finds progress fighting hate speech’, *Meta Newsroom*, 23 June 2020, <https://about.fb.com/news/2020/06/progress-fighting-hate-speech/>.

Misinformation

The pandemic has increased concern about misinformation and whether digital platforms are creating echo chambers. The factor that is often cited is concern around “algorithms”. Below we provide more detail about the role of algorithms and the integrity measures we take to reduce the distribution of problematic content including misinformation. To assist in understanding how we remove and reduce the distribution of harmful misinformation, it is first helpful to understand our approach to this type of content.

We take a significant number of steps to combat harmful misinformation, especially in relation to misinformation on COVID-19. These fall under a three-part framework.

1. **Remove** misinformation that could cause imminent, physical harm; and we allow for appeals in instances where we may not get this right;
2. **Reduce** the spread of fact-checked misinformation; and
3. Promote authoritative information and develop tools to **inform** our users.

This approach is supplemented by research and partnerships to ensure that we are constantly updating our efforts in response to changes in the nature of misinformation.

We are committed to working with policymakers and partners around the world to combat misinformation, and we have worked constructively with Government and industry in Australia to increase accountability and transparency around our misinformation efforts.

In particular, in 2020 we became a founding member and signatory to the Australian Disinformation and Misinformation Industry Code.⁸⁶ The code is a major step in establishing a regulatory framework around industry’s work to combat misinformation and disinformation, with other countries around the world looking to emulate this approach. We provide further detail on the code below.

Meta has committed to 43 specific commitments to meet the obligations outlined in the voluntary code, and has begun reporting annually on our commitments (beginning with our first annual report in May 2021).⁸⁷ This provides greater transparency to Australian

⁸⁶ J Machin, ‘Facebook’s response to Australia’s disinformation and misinformation industry code’, *Facebook Australia Blog*, 21 May 2021, <https://australia.fb.com/post/facebook-response-to-australias-disinformation-and-misinformation-industry-code/>

⁸⁷ Further detail on Facebook’s commitments can be found at <https://australia.fb.com/post/facebook-response-to-australias-disinformation-and-misinformation-industry-code/>

policymakers and the community about the steps we take to combat disinformation and misinformation.

Importantly, as part of the code, we released Australia-specific statistics about content on our platform. Country-specific statistics about online content have limitations: they do not provide the full picture of what content Australians might see online. However, we recognise the importance of providing some data to contribute to a sophisticated public policy debate about misinformation in Australia, while industry and experts consider the best ways to measure and report on phenomena like online misinformation in the long term.

- At the time of publishing our response to the Misinformation Code in May 2021, we reported that since the beginning of the pandemic in March 2020 to the end of December 2020, globally Meta had removed over 14 million pieces of content that constituted misinformation related to COVID-19 that may lead to harm, such as content relating to fake preventative measures or exaggerated cures. Of these, 110,000 pieces of content were removed from Australia (noting that Australians benefitted from the content we removed from other countries as well).

Since these figures were released, we have updated our global statistics on misinformation. In August 2021, we reported that since the beginning of the pandemic, we have removed more than 20 million pieces of content from Facebook and Instagram globally for violating our policies on COVID-19 related misinformation. We have also displayed warning labels on more than 190 million pieces of COVID-related content that our third-party fact checking-partners rated as false, partly false, altered or missing context, to limit the spread of COVID-19 and vaccine misinformation.⁸⁸

- We have made a COVID-19 Information Centre available around the world to promote authoritative information to Facebook users. More than 2 billion people globally, including over 6.2 million people in Australia visited the COVID-19 Information Centre during 2020.⁸⁹

Our approach to misinformation is outlined in more detail below.

⁸⁸ Meta, *Community Standards Enforcement Report Q2 2021*, <https://transparency.fb.com/data/community-standards-enforcement/>

⁸⁹ J Machin, 'Facebook's response to Australia's disinformation and misinformation industry code', *Facebook Australia Blog*, 21 May 2021, <https://australia.fb.com/post/facebooks-response-to-australias-disinformation-and-misinformation-industry-code/>

Remove

Meta removes misinformation that violates our Community Standards and can cause imminent, physical harm. These policies are continuously updated to reflect the latest research, and keep pace with changes happening online and offline around the world.

We define misinformation as claims that are misleading or false, and we will remove any content that violates our policies that relate to misinformation, these include:

- **Misinformation & Harm policy.** We have had a policy on Misinformation and Harm since 2018.

We work with experts around the world - in particular, the World Health Organization - to identify COVID-related claims that could cause imminent, physical harm. In December 2020, we expanded our policy to cover false claims about COVID vaccines, and in January 2021 we expanded our Misinformation and Harm policy to include claims about vaccines generally.⁹⁰ In October 2021, we expanded our policy to include a range of false claims about COVID-19 vaccines and children.⁹¹

- **Election-related misinformation that may constitute voter fraud and/or interference.** Under our policies, we prohibit misrepresentation of the dates, locations, times, and methods of voting or voter registration (for example: claims that you can vote using an online app); misrepresentations of who can vote, how to vote, qualifications for voting and whether a vote will be counted; or misrepresentation of who can vote, qualifications for voting, whether a vote will be counted, and what information or materials must be provided in order to vote. We also do not allow statements that advocate, provide instructions, or show explicit intent to illegally participate in a voting process.

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.

⁹⁰ K Jin, 'Keeping people safe and informed about the coronavirus', *Meta Newsroom*, 18 December 2020, <https://about.fb.com/news/2020/12/coronavirus/#removing-covid-vaccine-misinformation>; G Rosen, An update on our work to keep people informed and limit misinformation about COVID-19, *Meta Newsroom*, 16 April 2020, <https://about.fb.com/news/2020/04/covid-19-misinfo-update/#removing-more-false-claims>

⁹¹ K Jin, 'Supporting covid-19 vaccination efforts for children', *Meta Newsroom*, 29 October 2021, <https://about.fb.com/news/2021/10/supporting-covid-19-vaccine-children/>

- **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 Facebook profiles and 40,800 Instagram accounts. Some of these Pages, groups, events, profiles and accounts were located in Australia.⁹²
- **Manipulated media, also known as “deepfakes”, in line with our Manipulated Media policy.** After consulting with more than 50 global experts with technical, policy, media, legal, civic and academic backgrounds, we announced in 2020 that we would remove manipulated media if: (1) it has been edited or synthesised – beyond adjustments for clarity or quality – in ways that aren’t apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say; and (2) it is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic.⁹³

Reduce

For content that does not violate our Community Standards but is still problematic or otherwise low-quality, we reduce its distribution. Our initiatives to reduce the distribution of misinformation include:

- **Third-party fact-checking program.** 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. In Australia, we partner with the Australian Associated Press and Agence France Presse, both certified by the nonpartisan International FactChecking Network. All fact-checks by these partners are publicly available on their websites.⁹⁴

⁹² Meta, ‘An update to how we address movements and organizations tied to violence’, *Meta Newsroom*, updated 9 November 2021, <https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/>

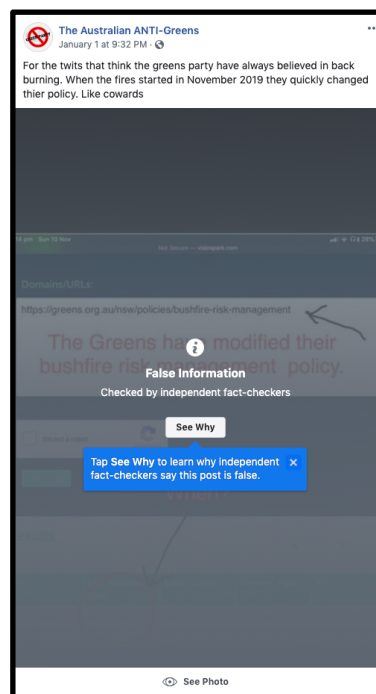
⁹³ M Bickert, Enforcing Against Manipulated Media, *Meta Newsroom*, 6 January 2020, <https://about.fb.com/news/2020/01/enforcingagainst-manipulated-media/>

⁹⁴ 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>

Since 2016, Meta has contributed more than \$84 million globally to support our fact-checking efforts.⁹⁵ This includes direct support of fact-checkers for their work on our platforms, as well as industry initiatives like sponsorships, fellowships, and grant programs. We now work with over 80 fact-checking partners around the world covering more than 60 languages.

- **Warning labels.** Once a third-party fact-checking partner rates a post as ‘false’, we apply a warning label or notice and show a debunking article from the fact checker, shown in Figure 12 below. It is not possible to see the content without clicking past the warning label. When people see these warning labels, 95 per cent of the time they do not go on to view the original content.

Figure 12: Misinformation warning label



- **Ensuring that fewer people see false information.** Once a fact-checker rates a piece of content as “false”, “partly false” or “altered”, it appears lower in News Feed on Facebook. On Instagram, it gets filtered out of Explore and is featured less prominently in feed and stories. This significantly reduces the number of

⁹⁵ C Alexander, ‘Facebook launches accelerator challenge for global fact-checkers to expand reach of reliable information’, *Facebook Journalism Project*, 26 August 2021, <https://www.facebook.com/journalismproject/accelerator-fact-checkers>

people who see it. We also reject ads with content that has been rated by fact-checkers.

- **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.⁹⁶
- **Taking action on Pages, Groups, accounts, or websites found to repeatedly share misinformation, including removing them from recommendations.** When Pages, Groups 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're about to join a group that has Community Standards violations, so they can make a more informed decision before joining.⁹⁷ If they continue to share misinformation, the Page or Group is removed in its entirety. This includes Pages operated by public figures.

Inform

Meta supports Government's initiatives by connecting people with up-to-date, authoritative information, these include:

- **Promoting authoritative information.** As outlined above, we've worked with the World Health Organization to create a COVID-19 Information Centre with verified, authoritative information about COVID-19.

In November 2021, we also launched the Climate Science Information Centre in Australia.⁹⁸ The Centre will connect users to authoritative information from leading climate organisations around the world.

We have also announced a US\$1 million for a grant program with the International Fact Checking Network, to support organisations to combat climate

⁹⁶ G Rosen, An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19, *Meta Newsroom*, updated 12 May 2020, <https://about.fb.com/news/2020/04/covid-19-misinfo-update/>,

⁹⁷ T Alison, 'Changes to keep Facebook Groups safe', *Meta Newsroom*, 17 March 2021, <https://about.fb.com/news/2021/03/changes-to-keep-facebook-groups-safe/>

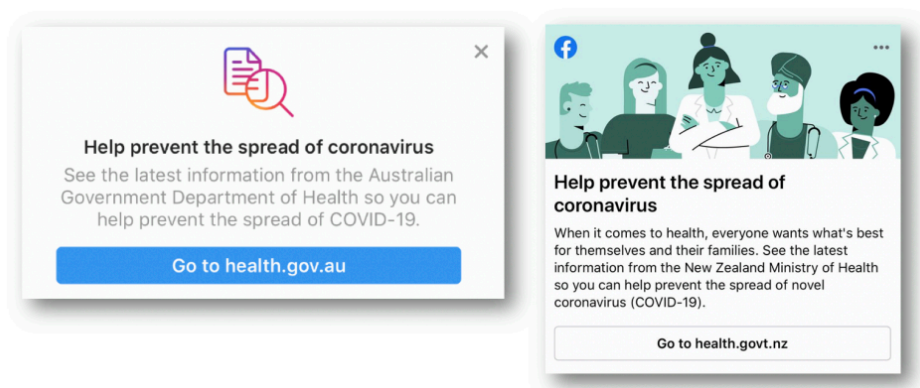
⁹⁸ N Clegg, 'Our commitment to combatting climate change', *Meta Newsroom*, 1 November 2021, <https://about.fb.com/news/2021/11/our-commitment-to-combating-climate-change/>

misinformation.⁹⁹

- **Providing free advertising credits to Government departments.** To support the rollout of the COVID-19 vaccine in Australia, Meta has given millions of dollars worth of Facebook advertising credits to the Australian Federal and State Governments. This helps to ensure the promotion of authoritative vaccine information from governments around the country.
- **Using in-product tools to help authorities communicate time critical information on key COVID-19 related topics.** In June 2021, we announced the expansion of our Local Alerts product.¹⁰⁰ This tool allows Departments and agencies to communicate directly with Facebook users on time critical announcements such as COVID outbreaks or lockdown situations. So far, Australia is the only country outside of the US to get access to this important product.

We also display prompts, shown in Figure 13, on Facebook and Instagram to direct users to official sources of information, including the Australian Government and the World Health Organization. These have been seen by every Facebook and Instagram user in Australia multiple times, either in their Feeds or when they search for coronavirus-related terms. We also ran prompts in Australia to urge people to wear a mask.

Figure 13: COVID-19 in-product prompts



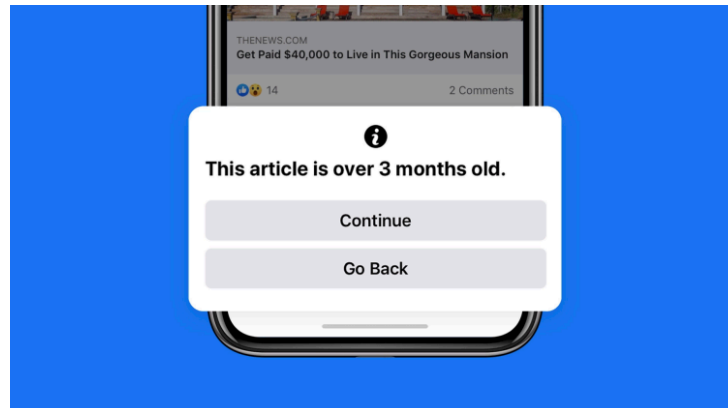
- **Providing contextual information around posts that users see from public Pages.** We have developed a number of other labels and signals to indicate the

⁹⁹ Meta, 'Tackling climate change together', *Meta Newsroom*, 16 September 2021, <https://about.fb.com/news/2021/09/tackling-climate-change-together/>

¹⁰⁰ J Machin, 'Facebook expands Local Alerts Tool', *Facebook Australia Blog*, 29 June 2021, <https://australia.fb.com/post/facebook-expands-local-alerts-tool/>

trustworthiness of posts they see on Facebook. These include the context button, which provides information about the sources of articles in News Feed;¹⁰¹ the breaking news tag, to help people easily identify timely news or urgent stories;¹⁰² and a notification screen that lets people know when news articles they are about to share are more than 90 days old, shown in Figure 14 below.¹⁰³

Figure 14: Contextual Information Label



Partnerships and research

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

- In September 2021, under our Australia-specific partnership with misinformation experts First Draft, we launched a “Don’t Be a Misinfluencer” campaign for public figures and creators, shown in Figure 15. The campaign aims to prevent the amplification of misinformation and includes a toolkit with information on how to identify and combat misinformation.¹⁰⁴

¹⁰¹ J Smith, A Leavitt & G Jackson, ‘Designing New Ways to Give Context to Stories’, *Meta Newsroom*, 8 April 2018, <https://about.fb.com/news/2018/04/inside-feed-article-context/>

¹⁰² J Rhyu, ‘Enabling Publishers to Label Breaking News on Facebook’, *Facebook for Journalism*, 6 April 2020, <https://www.facebook.com/journalismproject/facebook-breaking-news-label>

¹⁰³ J Hegeman, ‘Providing people with additional context about content that they share’, *Meta Newsroom*, 25 June 2020, <https://about.fb.com/news/2020/06/more-context-for-news-articles-and-other-content/>

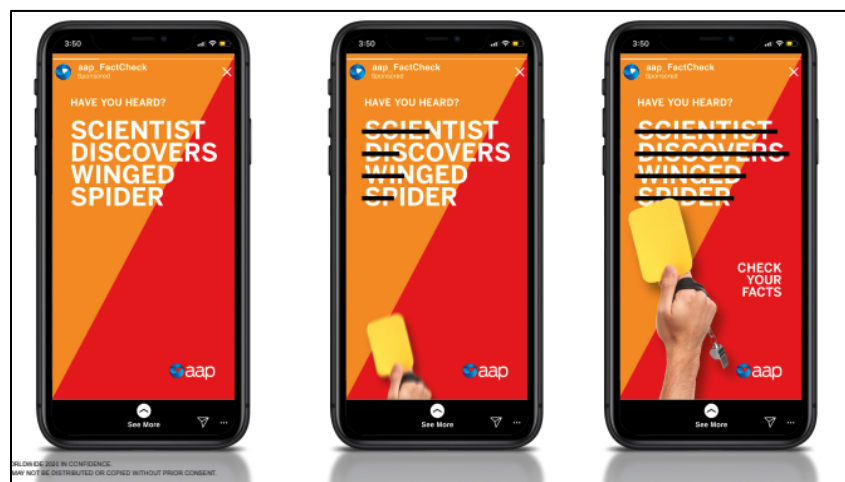
¹⁰⁴ First Draft, ‘Protect your voice: a toolkit for Australian influencers and celebrities’, *First Draft website*, <https://firstdraftnews.org/tackling/protect-your-voice-a-toolkit-for-australian-influencers-and-celebrities/>

Figure 15: First Draft ‘Don’t Be A Mis-Influencer’ Campaign



- In October 2021, we launched a media literacy ‘Check the Facts’ initiative for Australians with the Australian Associated Press, shown in Figure 16. The campaign uses videos and social tiles to teach Australians about the importance of fact-checking, and how to recognise and avoid the spread of misinformation.¹⁰⁵

Figure 16: Australian Associated Press ‘Check the Facts’ Campaign



¹⁰⁵ Australian Associated Press, ‘Australians urged to Check the Facts’, *AAP website*, 25 October 2021, <https://www.aap.com.au/news/australians-urged-to-check-the-facts/>

- We sponsored the Business Council of Australia in their ‘One Shot Closer Campaign’.¹⁰⁶ The campaign brings together employers across Australia to speak in a united voice to boost vaccination rates across Australia.
- Meta has joined UNICEF Australia’s Vaccine Alliance that encourages cross-sector collaboration in Australia and around the world to provide equitable access to the vaccine.¹⁰⁷ Meta has contributed free advertising to the alliance to promote and encourage people to get vaccinated.
- We work with Australian Government, state and territory agencies, to refer content for review and potential removal if they violate our Community Standards.

We are also undertaking a number of pieces of research, globally and Australia-specific, to understand the phenomenon of misinformation. This includes:

- Commissioning independent research by Australian academic Dr Andrea Carson to map government approaches to combatting misinformation around the world, focussing on the Asia-Pacific region¹⁰⁸. The report ‘Tackling Fake News’ was launched in January 2021 and has been positively received globally by policymakers and experts as they consider new approaches to misinformation.
- Supporting the Australian Media Literacy Alliance to conduct the first Australian national media literacy survey. The results of the survey were released in October 2021 alongside recommendations for governments, companies and communities to improve media literacy.¹⁰⁹
- Meta supported an analytical paper by First Draft on disinformation and misinformation amongst diaspora groups with a focus on Chinese language.¹¹⁰ The paper aims to inform policymakers on how to reduce misinformation within Chinese diaspora communities ahead of the next federal election.

¹⁰⁶ Business Council Australia, ‘One shot closer’ Campaign, <https://www.oneshotcloser.com.au/>

¹⁰⁷ UNICEF Australia, ‘COVID vaccination alliance’, <https://www.unicef.org.au/about-us/partnerships/covid-vaccination-alliance>

¹⁰⁸ A Carson, ‘Fighting Fake News: A Study of Online Misinformation Regulation in the Asia-Pacific’, https://www.latrobe.edu.au/_data/assets/pdf_file/0019/1203553/carson-fake-news.pdf

¹⁰⁹ Australian Media Literacy Alliance, ‘Towards a national strategy for media literacy’, October 2021, https://medialiteracy.org.au/wp-content/uploads/2021/10/AMLA-Consultation-Workshop-Report_UPDATE-25-10-2021.pdf

¹¹⁰ E Chan, S Zhang, ‘Disinformation, stigma and chinese diaspora: policy guidance for Australia’, *First Draft website*, 31 August 2021, <https://firstdraftnews.org/long-form-article/disinformation-stigma-and-chinese-diaspora-policy-guidance-for-australia/>

Algorithms

In recent years, concern has been expressed by some commentators that social media is fuelling polarisation, exploiting human weaknesses and insecurities, or creating echo chambers where everyone gets their own slice of reality, eroding the public sphere and undermining the understanding of common facts. The factor that is often cited is concern around “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. For example, if you can set your heat to turn on at a given temperature or have the lights in your house on a timer, you are probably using an algorithm.

The subset of algorithms normally being referred to by commentators discussing social media are those that are used to rank and distribute content.

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 News Feed on Facebook. Those algorithms that help with ranking play different roles. Some help us find and remove content from our platform that violates our Community Standards. 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 News Feed or Instagram Feed 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’s why we have been releasing products, tools and greater transparency about the way algorithms work on our services. Our Content Distribution Guidelines and Recommendation Guidelines, 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.

As our Vice-President of External Affairs, Nick Clegg, has outlined, we want to continue working with policymakers to explain how algorithms work on our services and to help build confidence in the community about the role that algorithms play in their experience online.¹¹¹

How our ranking algorithms work

On Facebook and Instagram, one of the ways that people connect with friends, family and other accounts that they follow is via a “Feed” (called “News Feed” on Facebook).

Historically, these feeds showed content in chronological order, as more people started using our services and more content was shared. However, 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.¹¹² Similarly, on Facebook, the goal of News Feed is to arrange the posts from friends, Groups and Pages you follow to show you what matters most to you at the top of your feed. Our ranking algorithms use thousands of signals to rank posts for each person’s News Feed with this goal in mind.¹¹³ As a result, each person’s News 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 to them, every time they come to Facebook or Instagram.

When it comes to Facebook, the average person has thousands of posts they potentially could see at any given time, so to help them find the content we hope they will find most meaningful, we use the ranking process, which orders the posts in News Feed, putting the things we think you will find most valuable closest to the top. The idea is that this results in content from your best friend being placed high in your Feed, while content from an acquaintance you met several years ago will often be much lower down.

Every piece of content that could potentially feature in a person’s News Feed — including the posts someone has not seen from their “friends,” the Pages they follow, and Groups they have joined — goes through the ranking process. We call that universe of content someone’s inventory. Because we have billions of people using our services and

¹¹¹ N Clegg, You and the algorithm: It takes two to tango, 31 March 2021, <https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2>

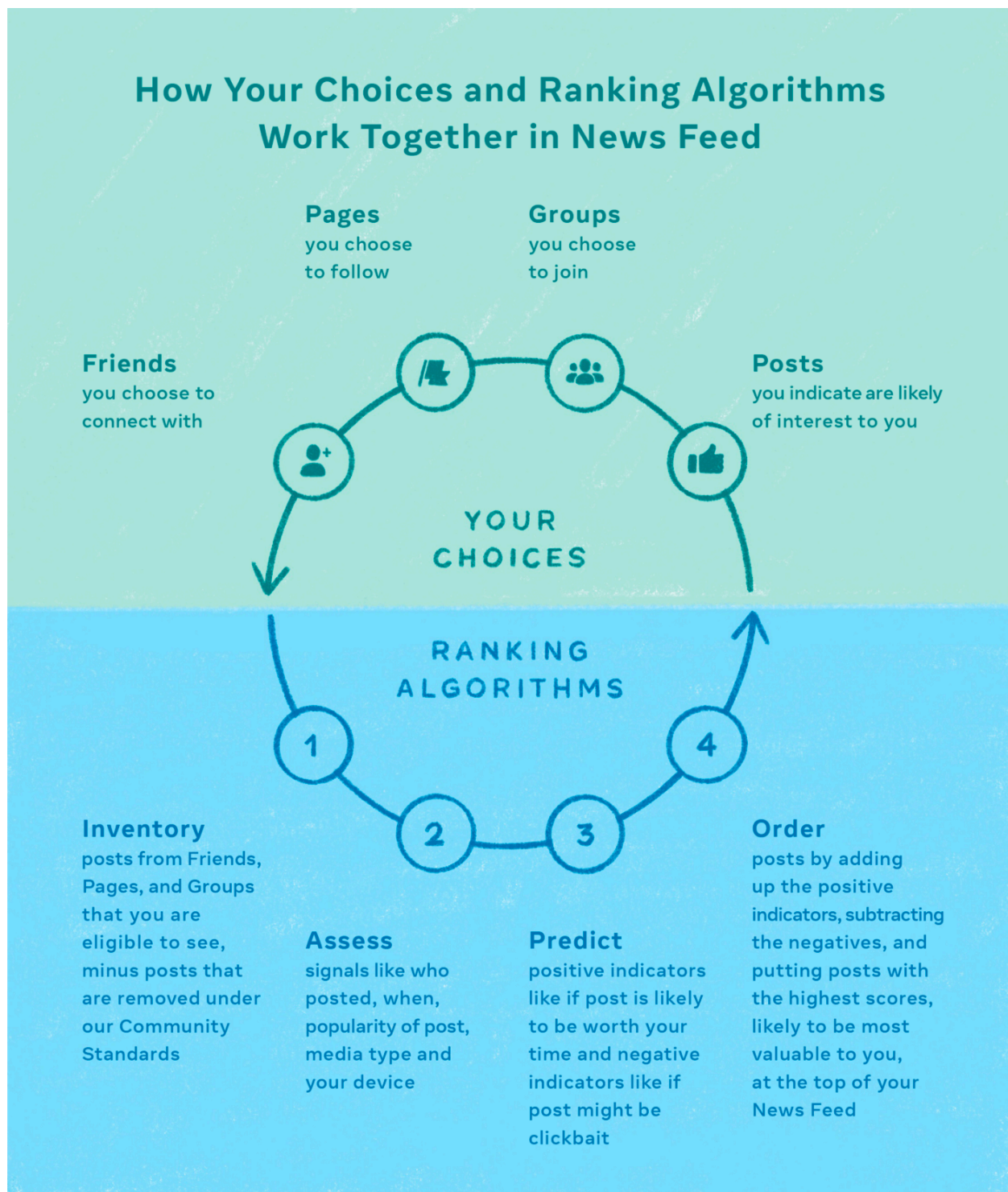
¹¹² A Mosseri, Shedding more light on how Instagram works, *Instagram Blog*, 8 June 2021, <https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works>

¹¹³ A Lada, M Wang, How does News Feed predict what you want to see? *Meta Newsroom*, 26 January 2021, <https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/>

thousands of pieces of content that could potentially be seen in News Feed for most of them, 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's a photo, video or link, how popular it is on the platform, or the type of device you are using, see Figure 17 below for more detail. 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 “like” 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. You can think about this sort of like a spam filter in your inbox: it helps filter out content you won't find meaningful or relevant, and prioritises content you will. This can be thought of as a kind of a spam filter in your inbox: it helps filter out content people will not find meaningful or relevant, and prioritises content that they will.

Figure 17: How Your Choices and Ranking Algorithms Working Together in News Feed



One way we measure whether something creates long-term value for a person is to ask them. For example, we survey people¹¹⁴ to ask how meaningful they found an interaction or whether a post was worth their time, so that our system reflects what people enjoy and find meaningful.¹¹⁵ 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 how often people like it, comment on it, or share it — can be a helpful indicator that it’s 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 News Feed experience. We are currently working on building out these surveys by asking new questions about the content people find valuable, and we recently made it much easier for people to tell us what content they do not enjoy seeing in their News Feed.¹¹⁶

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, and posts deemed false by one of the more than 80 independent fact checking organisations that evaluate Facebook content. We recently published a list of all of the types of problematic content and behaviour that receive reduced distribution on News 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 you 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 your Feed.

¹¹⁴ R Sethuraman, Using surveys to make News Feed more personal, *Meta Newsroom*, 16 May 2019, <https://about.fb.com/news/2019/05/more-personalized-experiences/>

¹¹⁵ Meta, How users help shape Facebook, *Meta Newsroom*, 13 July 2018, <https://about.fb.com/news/2018/07/how-users-help-shape-facebook/>

¹¹⁶ A Gupta, Incorporating more feedback into News Feed ranking, *Meta Newsroom*, 22 April 2021, <https://about.fb.com/news/2021/04/incorporating-more-feedback-into-news-feed-ranking/>

Tools for control and transparency

We understand concerns around the lack of transparency over how algorithmic ranking systems work, so we have introduced new measures to give people more insight into and control over how content appears in their Feed. We also provide regular updates about changes, new features or feedback tools, and the guidelines we use to rank content.

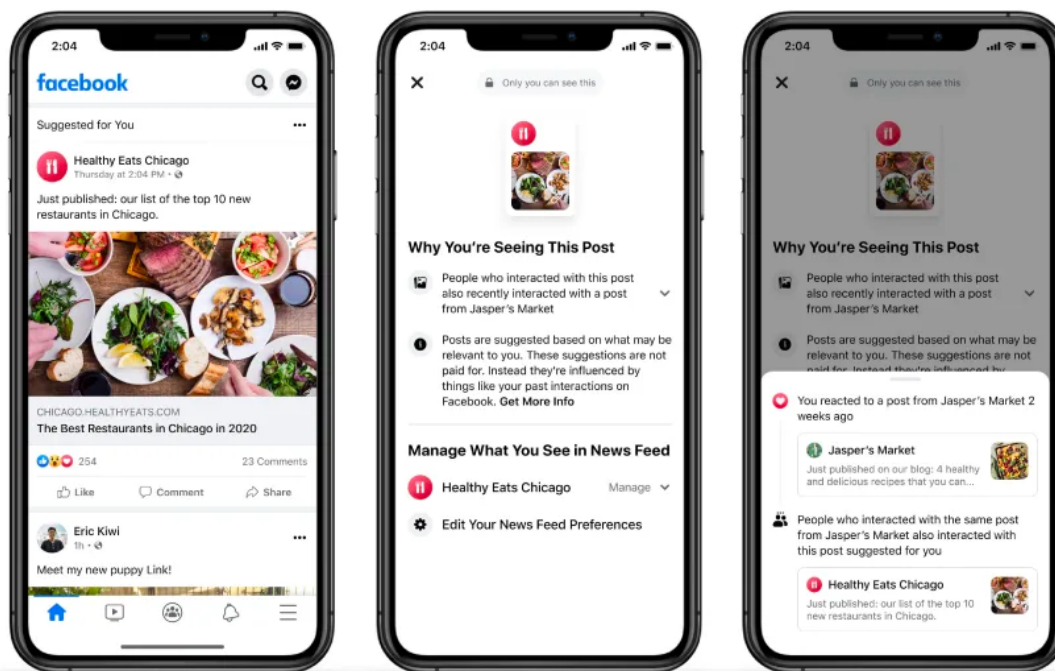
Some of the tools that provide people with greater insight and control over their experience include:

- **Why am I seeing this post?** This feature, shown in Figure 18, was launched in March 2019¹¹⁷ to help people understand and more easily control what they are seeing in their News Feed. This tool explains how a person's past interactions impact the ranking of posts in their Feed. For example, if the post is from a friend, a Group you joined, or a Page you followed, the information generally that has the largest influence over the order of posts, including: (a) how often you interact with posts from people, Pages or Groups; (b) how often you interact with a specific type of post, for example, videos, photos or links; and (c) the popularity of the posts shared by the people, Pages and Groups you follow. This was recently expanded to include context about why people are seeing suggested posts.¹¹⁸

¹¹⁷ Meta, Why Am I Seeing This? We have an answer for you, *Meta Newsroom*, 31 March 2019, <https://about.fb.com/news/2019/03/why-am-i-seeing-this/>

¹¹⁸ R Sethuraman, More control and context in News Feed, *Meta Newsroom*, 31 March 2021, <https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/>

Figure 18: Why am I seeing this post?



- **Why am I seeing this ad?** This feature allows people to see how factors like basic demographic details, interests and website visits contribute to the ads in their Facebook Feed. There are also additional details about when information on an advertiser's list matches a person's Facebook profile.¹¹⁹
- **News Feed Preferences.** This is a suite of tools that allow people to manage what they see in their Facebook Feed, including the ability to unfollow people, snooze a particular account, or prioritise Favourites.¹²⁰
- **Favourites.** This tool allows Facebook users to control and prioritise posts from the friends and Pages they care about most. By selecting up to 30 friends and Pages to include in Favourites, their posts will appear higher in ranked News Feed and can also be viewed in a separate feed populated exclusively with posts from a person's "Favourites".¹²¹ We have also begun testing this option on Instagram.

¹¹⁹ Meta, Understand why you're seeing certain ads and how you can adjust your ad experience, *Meta Newsroom*, 11 July 2019, <https://about.fb.com/news/2019/07/understand-why-youre-seeing-ads/>

¹²⁰ Facebook, *How can I see and adjust my Facebook News Feed preferences?*, <https://www.facebook.com/help/371675846332829>

¹²¹ R Sethuraman, More control and context in News Feed, *Meta Newsroom*, 31 March 2021, <https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/>

- **Feed Filter Bar.** This feature allows Facebook users to alternative between News Feed experiences - the algorithmically-ranked News Feed, the chronological Most Recent feed¹²², or the Favorites Feed discussed above.¹²³ We have also recently announced that we will soon provide users with the option to have a chronological feed on Instagram.¹²⁴

Providing guidelines for ranking

In addition to providing on-platform tools to increase the transparency around why people see particular content or ads, we also provide transparency around ranking algorithms by publishing the News Feed Values, content ranking guidelines and details of any updates.

We often make improvements to News Feed, and when we do, we rely on a set of core values.¹²⁵ These values guide our thinking, and help us keep the central experience of News Feed intact as it evolves. In summary, the values are:

- Friends and family come first
- A platform for all ideas
- Authentic communication
- You control your experience
- Constant iteration.

As mentioned above, we also recently published Facebook's Content Distribution Guidelines to share more detail on the types of content that we demote in News Feed.¹²⁶ While the Community Standards make it clear what content is removed from our services because we don't allow it, the Content Distribution Guidelines make it clear what content receives reduced distribution on News Feed because it's 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.

¹²² Facebook, *How do I see the most recent posts in my News Feed on Facebook?*
<https://www.facebook.com/help/218728138156311>

¹²³ R Sethuraman, More control and context in News Feed, *Meta Newsroom*, 31 March 2021,
<https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/>

¹²⁴ Instagram Comms, <https://twitter.com/InstagramComms/status/1468707110036852750?s=20>

¹²⁵ A Mosseri, Building a better News Feed for you, *Meta Newsroom*, 29 June 2016,
<https://about.fb.com/news/2016/06/building-a-better-news-feed-for-you/>

¹²⁶ Meta, Types of content we demote, *Transparency Centre*, 20 December 2021,
<https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/>

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've 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 don't like seeing on Facebook and make changes to News Feed 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 for our community, regardless of the intent. We'll make this content more difficult for people to encounter.

We will continue to update the Content Distribution Guidelines. The Widely Viewed Content Report (that is discussed in the Transparency and Accountability section above) 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.

Finally, we continually evaluate the effectiveness of Feed ranking signals. We share updates about the biggest changes and tests we have launched on our Inside Feed blog to give people who use Facebook more control over their News Feed.¹²⁷ Beyond sharing information about specific ranking changes, we are also making an effort to provide people with more detail about our ranking processes in general. For example, last year, the CEO of Instagram published a blog post detailing the ranking process on Instagram from start to finish.¹²⁸

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

¹²⁷ Meta, *Inside Feed*, <https://about.fb.com/news/category/inside-feed/>

¹²⁸ 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>

our recommendations experiences include Pages You May Like, "suggested for you" posts in News Feed, People You May Know or Groups You Should Join.

Since recommended content doesn't come from accounts that people have already chosen to follow, it's 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.¹²⁹ In developing these guidelines, we consulted 50 leading experts specialising in recommendation systems, expression, safety and digital rights. Recommendation Guidelines are available for both Facebook¹³⁰ and Instagram.¹³¹

The impact of algorithms

Social media lets people discuss, share, and criticise freely and at scale, without the boundaries or mediation previously imposed by the gatekeepers of the traditional media industry. For hundreds of millions of people, it is the first time that they have been able to speak freely and be heard in this way, with no barrier to entry apart from an internet connection. People do not just have a video camera in their pocket — with social media, they also have the means to distribute what they see. This is a dramatic and historic democratisation of speech.

Central to many of the charges by critics of social media is the idea that algorithmic systems actively encourage the sharing of sensational content and are designed to keep people scrolling endlessly. Of course, on a platform built around people sharing things they are interested in or moved by, content that provokes strong emotions is invariably going to be shared. At one level, the fact that people respond to sensational content isn't new. As generations of newspaper sub-editors can attest, emotive language and arresting imagery grab people's attention and engage them. It's human nature. But

¹²⁹ G Rosen, 'Recommendation guidelines', *Meta Newsroom*, 31 August 2020, <https://about.fb.com/news/2020/08/recommendation-guidelines/>

¹³⁰ Facebook, 'What are recommendations on Facebook?', *Help Centre*, <https://www.facebook.com/help/1257205004624246>

¹³¹ Instagram, 'What are recommendations on Instagram?', *Help Centre*, <https://help.instagram.com/313829416281232>

Meta's systems are not designed to reward provocative content. In fact, key parts of those systems are designed to do just the opposite.

The reality is, that it is not in Meta's interest — financially or reputationally — to continually turn up the temperature and push users towards ever more extreme content. The company's long-term growth will be best served if people continue to use its products for years to come. If our company prioritised keeping people online an extra 10 or 20 minutes, but in doing so made them less likely to return in the future, it would be self-defeating.

Additionally, the vast majority of Facebook's revenue comes from advertising. Advertisers don't want their brands and products displayed next to extreme or hateful content.

The impact of algorithms is often discussed in relation to two issues: mental health; and political polarisation in society.

Mental health

As outlined in the 'Mental Health and Wellbeing' section above, our research and other academic literature suggests that it's about how you use social media that matters when it comes to your wellbeing. This ability to connect with relatives, classmates, and colleagues is what drew many of us to Facebook and Instagram in the first place, and so it is no surprise that staying in touch with these friends and loved ones brings us joy and strengthens our sense of community.

Over the years, we have worked to make our ranking algorithms more about social interaction and less about spending time; we made changes in 2018 to provide more opportunities for meaningful interactions - even if it decreased the amount of time people spent on our platform in the short term, we have reduced the distribution of things like clickbait headlines and false news, even though people often click on those links at a high rate, and we optimise ranking so posts from the friends you care about most are more likely to appear at the top of your feed. This is what people tell us in surveys that they want to see. Similarly, our ranking promotes posts that are personally informative. We have also redesigned the comments feature to foster better conversations.¹³²

¹³² S Nguyen and Freitas, 'Making News Feed an Easier Place to Connect and Navigate', *Meta Newsroom*, 15 August 2017, <https://about.fb.com/news/2017/08/making-news-feed-an-easier-place-to-connect-and-navigate/>

Polarisation

Political and social polarisation has been the subject of swathes of serious academic research in recent years — the results of which are in truth mixed, with many studies suggesting that social media is not the primary driver of polarisation after all, and that evidence of the ‘filter bubble’ effect is thin at best.

Many studies indicate polarisation has not been increasing in Australia and, in countries like the US where “affective polarisation” (the measure of someone’s negative feelings about the opposite party) has increased, research suggests that social media is not the primary driver of polarisation. More details about this are discussed in a recent Medium post by our Vice President of Global Affairs Nick Clegg.¹³³

One piece of research in particular covers Australia: Boxell, Gentzkow and Shapiro’s article on Cross-Country Trends in Affective Polarisation.¹³⁴ These researchers examined trends in affective polarisation across nine OECD countries over the past 40 years. The findings demonstrate that polarisation has remained stable, with a slight decrease in Australia since the mid-1990s.

Research from Stanford last year looked in depth at trends in nine countries over 40 years, and found that in some countries polarisation was on the rise before Facebook even existed, and in others it has been decreasing while internet and Facebook use increased.¹³⁵ A recent study found in one country (Bosnia and Herzegovina) during a polarising time (the anniversary of conflict in the former Yugoslavia), *deactivating* Facebook led to increased affective polarisation.¹³⁶ Other credible recent studies have found that polarisation in the United States has increased the most among the demographic groups least likely to use the internet and social media,¹³⁷ and data

¹³³ N Clegg, ‘You and the algorithm: It takes two to tango’, *Medium*, 31 March 2021, <https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2>

¹³⁴ L Boxell, M Gentzkow & J Shapiro, *Cross-Country Trends in Affective Polarization*, June 2021, <https://web.stanford.edu/~gentzkow/research/cross-polar.pdf>

¹³⁵ *ibid.*

¹³⁶ N Asimovic et al., ‘Testing the effects of Facebook usage in an ethnically polarised setting’, *Proceedings of the National Academy of Sciences of the United States of America*, 22 June 2021, <https://www.pnas.org/content/118/25/e2022819118>

¹³⁷ L Boxell, M Gentzkow, J Shapiro, ‘Greater Internet use is not associated with faster growth in political polarization among US demographic groups’, *Proceedings of the National Academy of Sciences of the United States of America*, 19 September 2017, <https://www.pnas.org/content/114/40/10612>

published in the EU suggests that levels of ideological polarisation are similar whether you get your news from social media or elsewhere.¹³⁸

A Harvard study ahead of the 2020 U.S. election found that election-related disinformation was primarily driven by elite, mass-media and cable news, and that social media played only a secondary role.¹³⁹ And research from both Pew¹⁴⁰ in 2019 and the Reuters Institute¹⁴¹ in 2017 showed that you're likely to encounter a more diverse set of opinions and ideas using social media than if you only engage with other types of media.

An earlier Stanford study¹⁴² showed that deactivating Facebook for four weeks before the 2018 US elections reduced polarisation on political issues, but also led to a reduction of people's news knowledge and attention to politics. However, it did not significantly lessen so-called "affective polarisation".

The available evidence simply does not support the idea that social media are the unambiguous driver of polarisation that many assert.

Whilst these existing research findings are helpful, we continue to invest in research. Over 2021 and 2022 we are investing over US\$4 million in a global round of funding for research on misinformation and polarisation. Four research proposals from Australian universities were granted funding for their work:

- 'Testing fact and logic-based responses to polarising climate misinformation' (John Cook and Sojung Kim, Monash University);
- 'How fact checkers compare: News trust and COVID-19 information' (Andrea Carson, James Meese, Justin B. Phillips, Leah Ruppanner, La Trobe University)¹⁴³;

¹³⁸ A Nguyen, H Tien Vu, 'Testing popular news discourse on the "echo chamber" effect: Does political polarisation occur among those relying on social media as their primary politics news source?'. *First Monday*, <http://eprints.bournemouth.ac.uk/32048/7/OSNs%20as%20a%20political%20news%20medium.pdf>

¹³⁹ Y Benkler et al., 'Mail-In Voter Fraud: Anatomy of a Disinformation Campaign', *Berkman Center Research Publication*, 8 October 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3703701

¹⁴⁰ L Silver, Christine Huang, 'In Emerging Economies, Smartphone and Social Media Users Have Broader Social Networks', *Pew Research Centre*, 22 August 2019, <https://www.pewresearch.org/internet/2019/08/22/in-emerging-economies-smartphone-and-social-media-users-have-broader-social-networks/>

¹⁴¹ N Neuman et al., Reuters Institute Digital News Report 2017, *Reuters Institute*, 2017, <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web%20.pdf>

¹⁴² H Allcot et al., 'The welfare effects of social media', *Stanford University*, 8 November 2019, <http://web.stanford.edu/~gentzkow/research/facebook.pdf>

¹⁴³ Meta, 'Announcing the 2021 recipients of research awards in misinformation and polarisation', *Meta Research*, 14 September 2021, <https://research.fb.com/blog/2020/08/announcing-the-winners-of-facebook-request-for-proposals-on-misinformation-and-polarization/>

- ‘Indigenous women and LGBTQI+ people and violence on Facebook’ (Bronwyn Carlson, Macquarie University); and
- ‘Unpacking trust and bias in social media news in developing countries’ (Denis Stukal, University of Sydney).¹⁴⁴

¹⁴⁴ A Leavitt, K Grant, ‘Announcing the winners of Facebook’s request for proposals on misinformation and polarization’, *Facebook Research*, 7 August 2020, <https://research.fb.com/blog/2020/08/announcing-the-winners-of-facebooks-request-for-proposals-on-misinformation-and-polarization/>

Transparency and accountability

We recognise that, as a large company, the decisions we take relating to content on our services can be significant. That's why we have supported a number of initiatives to ensure there is enhanced transparency and accountability for the decisions we take.

To provide the community, civil society and governments with greater confidence in how these decisions are made, Meta has been supportive of new regulations for content decisions. Facebook has been at the global forefront of calling for regulation of harmful content. In 2019, our CEO Mark Zuckerberg called for liberal democracies to develop new regulation in relation to online content (along with privacy, data portability and elections).¹⁴⁵

In addition, we released a white paper called 'Charting a Way Forward - Online Content Regulation', which raises a series of questions to assist in designing effective content regulation.¹⁴⁶

However, we are not waiting for regulation. We provide transparency about a wide range of areas – our community standards enforcement, government and law enforcement requests, content restrictions, internet disruptions, widely viewed content – among others.¹⁴⁷ And we have committed to external checks and audits of several of these transparency measures.

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

¹⁴⁵ M Zuckerberg, 'The Internet Needs New Rules', *Washington Post*, 30 March 2019, https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html

¹⁴⁶ M Bickert, *Charting a way forward: online content regulation*, white paper released February 2020, https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf.

¹⁴⁷ Meta, *Transparency Centre*, <https://transparency.fb.com/data/>

Transparency Reports

Maintaining transparency around how we make content decisions, and the nature and extent of the government requests we receive for user data, is really important to us.

We have a dedicated Transparency Centre¹⁴⁸ which gives our community visibility into how we enforce our policies, respond to data requests and protect intellectual property, while monitoring dynamics that limit access to Meta's platforms.

Community Standards Enforcement Report

Each quarter we release a Community Standards Enforcement Report (CSER). Our latest CSER was published in November 2021 and covers the period between July and September 2021.¹⁴⁹ We have cited the metrics from the most recent CSER throughout the submission.

The CSER is a voluntary transparency effort that allows for scrutiny of our efforts to enforce Facebook and Instagram's Community Standards, which outline what is and is not allowed on our services. Our CSER reports on five metrics:

1. **Content removed.** We measure the number of pieces of content (such as posts, photos, videos or comments) or accounts that we take action on for going against our standards. This metric shows the scale of our enforcement activity.
2. **Content removed proactively.** This metric shows the percentage of all content or accounts that we found and flagged before users reported them to us. We use this metric as an indicator of how effectively we detect violations.
3. **Prevalence.** Prevalence metrics allow us to track, both internally and externally, how much violating content people are seeing on our apps. Prevalence, in turn, helps us determine the right approaches to driving that metric down, whether it's through updating our policies, products or tools for our community.
4. **Appeals.** We report the number of pieces of content that people appeal after we take action on it for going against our policies. We offer appeals for the vast majority of violation types on Facebook and Instagram. We don't offer appeals for violations with extreme safety concerns, such as child exploitation imagery.

¹⁴⁸ Meta, *Transparency Centre*, <https://transparency.fb.com/data/>

¹⁴⁹ Meta, *Community Standards Enforcement Report* - <https://transparency.fb.com/data/community-standards-enforcement/>

Reporting on this metric holds us to account for our content decisions, and ensures we can continue to improve our enforcement.

5. **Restored content.** For policy violations, we measure the number of pieces of content (such as posts, photos, videos or comments) that we restored after we originally took action on them.

To ensure our CSER remains an accurate and meaningful measure of Meta's content moderation, we have appointed the Data Transparency Advisory Group (DTAG).¹⁵⁰ The DTAG is an independent body made up of international experts in measurement, statistics, criminology and governance who provide independent, public assessments of the CSER metrics, and the enforcement processes and measurement methodologies we use.

In 2019, the DTAG completed a formal review of the CSER.¹⁵¹ They found that Meta's approach to content moderation processes are appropriate given the scale at which we operate and the amount of content people post. They also found that the accuracy of our content review system was well designed, and the metrics we use to measure success (prevalence, actioned content and proactive rate), are in line with best practice.

DTAG also laid out 15 recommendations which Meta has continued to implement. These include additional metrics we should report on, further break-downs of the metrics we already provide, and making it easier for people who use Meta's services to stay updated on the changes we make to our policies.¹⁵²

In 2020, we committed to undergoing an independent audit to validate that our metrics published in the Community Standards Enforcement Report are measured and reported correctly. We have begun working with EY to conduct this audit, which is expected to be handed down in Q2 of 2022.¹⁵³

¹⁵⁰ Meta, 'An independent report on how we measure content and moderation', *Meta Newsroom*, 23 March 2019, <https://about.fb.com/news/2019/05/dtag-report/>

¹⁵¹ Ibid.

¹⁵² The full Report of the Facebook Data Transparency Advisory Group can be found here https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf

¹⁵³ V Sarang, 'Independent audit of Community Standards Enforcement Report metrics', *Meta Newsroom*, 11 August 2020, <https://about.fb.com/news/2020/08/independent-audit-of-enforcement-report-metrics/>.

Widely Viewed Content Report

In August 2021, we released the first 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 News Feed in the United States.¹⁵⁴ The report also includes additional insights into the different content types that appear, *e.g.*, posts with links, posts from groups, etc., to help people better understand Facebook's distribution systems and how those systems influence 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 have now released two WVCR's, which find that content that's *seen* by the most people isn't necessarily the content that also gets the most *engagement*. It also finds that the majority of content that people see on their News Feed are from friends and family, in line with News Feed changes we made in the past. The majority (87.1 per cent) of posts people see are from their family and friends or Groups and Pages that people follow. The most widely viewed posts in those quarters were from Pages focussed on sharing content about people's favourite movies, entertainment, cooking and family.¹⁵⁵

Since releasing our inaugural WVCR, we have engaged with academics, civil society groups and researchers to identify the parts of our first report they found most valuable, which metrics needed more context and how we can best support their understanding of content distribution on Facebook. Based on these discussions, we've provided more clarity into our methodology and included more context in the Companion Guide. Moving forward, we'll continue to work with external stakeholders to refine and improve these reports.

Other transparency reports

Over the years, we've expanded our Transparency Reports to give our community more visibility on how we action content on our platform. We regularly report on:

- **Government requests for user data.**¹⁵⁶ Meta responds to government requests for data in accordance with applicable law and our terms of service. Each request we receive is carefully reviewed for legal sufficiency and sufficient detail. Meta

¹⁵⁴ A Stepanov, 'Introducing the Widely Viewed Content Report', 18 August 2021, *Meta Newsroom*, <https://about.fb.com/news/2021/08/widely-viewed-content-report/>

¹⁵⁵ A Stepanov, 'Widely Views Content Report, Third Quarter 2021', *Meta Newsroom*, 9 November 2021, <https://about.fb.com/news/2021/11/facebook-widely-viewed-content-report-q3-2021/>

¹⁵⁶ Meta, *Government Requests for User Data*, <https://transparency.fb.com/data/government-data-requests/>

regularly produces this report on government requests for user data to provide information on the nature and extent of these requests and the strict policies and processes we have in place to handle them.

- **Content restrictions.**¹⁵⁷ We receive reports on content from governments and courts, as well from non-government entities. When content is reported as violating local law, but doesn't go against our Community Standards, we may limit access to that content in the country where the local violation is alleged. This report details instances where we limited access to content based on local law.
- **Internet disruptions.**¹⁵⁸ We oppose shutdowns, throttling and other disruptions of internet connectivity and are deeply concerned by the trend towards this approach in some countries. Even temporary disruptions of internet services can undermine human rights and economic activity. That's why we report the number of deliberate internet disruptions caused by governments around the world that impact the availability of our products.
- **Intellectual property report.**¹⁵⁹ We are committed to helping people and organisations protect their IP rights. We do not allow people to post content that violates someone else's IP rights. This report details how many reports of IP violations we received and how much content we took down on as a result.
- **Adversarial Threat Report.**¹⁶⁰ Each month we publish a list of coordinated inauthentic behaviour networks that we have taken down. In some cases, we share information about the action taken at the time of enforcement. Since 2017, our security teams at Facebook have identified and removed over 150 covert influence operations for violating our policy against CIB.

In addition to these transparency reports, we also provide data research tools – the Ad Library (a comprehensive, searchable database of ads running on Facebook and Instagram)¹⁶¹, Crowdtangle (a public insights tool about what's happening across

¹⁵⁷ Meta, *Content Restrictions*, <https://transparency.fb.com/data/content-restrictions/>

¹⁵⁸ Meta, *Internet Disruptions*, <https://transparency.fb.com/data/internet-disruptions/?from=https%3A%2F%2Ftransparency.facebook.com%2Finternet-disruptions>

¹⁵⁹ Meta, *Intellectual Property*, <https://transparency.fb.com/data/intellectual-property/>

¹⁶⁰ N Gleicher, 'Meta's Adversarial Threat Report', *Meta Newsroom*, 1 December 2021, <https://about.fb.com/news/2021/12/metas-adversarial-threat-report/>

¹⁶¹ Facebook, *Facebook Ad Library*, https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=US&media_type=all

Facebook, Instagram and Reddit)¹⁶² and the Facebook Open Research and Transparency (analysis tools and data for researchers to study the company's impact on the world in a privacy-protective way).¹⁶³

Independent Oversight

Oversight Board

To ensure greater accountability for our content governance, we have also taken proactive, voluntary steps to establish an Oversight Board. The Oversight Board makes binding rulings on difficult and significant decisions about content on Facebook and Instagram.

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.

As mentioned above, the Oversight Board comprises 40 experts in human rights and technology - including the Australian academic Professor Nic Suzor from Queensland University of Technology. The Board is entirely independent and hears appeals on Facebook's decisions relating to content on Facebook and Instagram. We have agreed that the Board's decisions will be binding, and the Board is also able to make recommendations about Facebook's policies.¹⁶⁴

The Oversight Board began issuing decisions in January 2021.¹⁶⁵ This includes a decision and policy recommendation related to our COVID-19 misinformation and harm policies.

¹⁶² Crowdtangle, <https://www.crowdtangle.com/>

¹⁶³ Facebook, *Facebook Open Research and Transparency*, <https://fort.fb.com/>

¹⁶⁴ B Harris, 'Establishing structure and governance for an independent oversight board', *Meta Newsroom*, 17 September 2019, <https://about.fb.com/news/2019/09/oversight-board-structure/>

¹⁶⁵ N Clegg, 'Welcome the oversight board', *Meta Newsroom*, 6 May 2020, <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/>

Since publishing its first decisions in January 2021, the Oversight Board has issued 18 decisions, and made more than 75 recommendations to Meta for future improvements.¹⁶⁶

The Oversight Board has also recently begun publishing Transparency Reports which provide new details on the Oversight Board's cases, decisions and recommendations. They will continue to be released each quarter. 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. These quarterly updates 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. You can find the Oversight Board Transparency Centre here <https://transparency.fb.com/en-gb/oversight/>

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

¹⁶⁶ Oversight Board, 'Oversight Board demands more transparency from Facebook', *Oversight Board*, October 2021, <https://oversightboard.com/news/215139350722703-oversight-board-demands-more-transparency-from-facebook/>

Privacy and data security

Meta's success is dependent on ensuring digital trust, and privacy and protecting data are at the heart of this.

Our approach to privacy provides users with meaningful transparency and control over how their data is used. We believe consumers need to be informed and empowered about the privacy choices available to them. To do this, we have worked to provide more user-friendly practices like layered privacy policies, just-in-time notices, and in-context notifications.¹⁶⁷ Privacy policies should not be the only ways that companies communicate with people about their information.

We've worked with policymakers, regulators, academics, civil society, businesses and other stakeholders over the years to develop approaches to data governance practices, the development of privacy enhancing technologies, and tools that show users how their information is used, and to allow them to manage it.¹⁶⁸ Our products aim to be transparent and informative so that people can easily access specific information about how we collect, use and share their personal information.

In addition to the service-wide privacy tools we put in place for our users, we also implement a number of specific measures to provide age-appropriate and privacy-protective experiences for young people. These controls put a number of default protections in place for those under the age of 18. They also help to empower young people to make the right choices about their experience online, and the information they want to see and share.

We provide some information about our own processes below, in the interests of openness and contributing to the public policy debate about how companies should approach privacy and protection of their users' data.

Internal data governance

We have created dozens of teams, both technical and non-technical, that are solely focused on putting privacy at the core of everything we do. Our privacy program is a company-wide and cross functional effort, with a centralised privacy function staffed by

¹⁶⁷ E Egan, *Communicating About Privacy: Towards People-Centred and Accountable Design*, white paper, <https://about.fb.com/wp-content/uploads/2020/07/Privacy-Transparency-White-Paper.pdf>

¹⁶⁸ Ibid.

hundreds of experts in privacy and data protection that is responsible for providing frameworks, tools, and infrastructure to preserve people's privacy.

As part of our governance, we have created a Privacy Committee – an independent committee of Meta's Board of Directors – that meets quarterly to ensure we live up to our privacy commitments. The Committee is made up of independent directors with a wealth of experience serving in similar oversight roles.

Privacy in our products

Our revamped Privacy Review process helps ensure every new product or feature is built with privacy in mind and provides people with choices and transparency. Privacy Review is a collaborative, cross-functional process used to address potential privacy risks and protect people's information, as we launch new and modified products and services.¹⁶⁹

During this review, cross-functional teams evaluate privacy risks associated with a project, and determine if there are any changes that need to happen before launch to control for those risks. This review considers whether a project meets our privacy expectations which include: purpose limitation, data minimisation, data retention, external data misuse, transparency and control, data access and management, fairness and accountability.¹⁷⁰

If there's no agreement between the members of the cross-functional team on what needs to happen, the team escalates to a central leadership review, and further to the CEO, if needed for resolution.

Privacy tools

We seek to build every product to be transparent so that people can understand how we collect, use and share data on demand. We also focus on communicating important information proactively, clearly and contextually.

Meta also maintains a public help center where anyone who visits the site can learn about privacy, safety, policies, reporting, and how to use our services, as well as our Data Policy

¹⁶⁹ M Protti, 'Sharing Progress on Our Privacy Work', *Meta Newsroom*, 23 October 2020, <https://about.fb.com/news/2020/10/sharing-progress-on-our-privacy-work/>.

¹⁷⁰ Meta, *Privacy Progress Update*, https://about.facebook.com/privacy-progress?_ga=2.17599677.564851063.1641853492-1552981365.1641853492

page for an explanation of the information we process to support Facebook, Instagram, Messenger.¹⁷¹

We have also announced the launch of a new Privacy Centre where users can learn about their privacy options and understand how we collect and user information. The Privacy Centre is being trialed initially, and will roll out to more people and apps in the coming months.¹⁷²

Our commitment to privacy is evident through the number of tools to give people transparency and control over how their data is used, including:

- **Manage Activity.** In June 2020, we launched the ‘Manage Activity’ tool which makes it easier for our users to manage their digital footprint on Facebook.¹⁷³ Manage Activity puts in one place the functions users need to search their activity and archive or delete as they choose. We built this tool to make it easy for users to curate their presence on Facebook to enable them to choose how to accurately reflect who they are.
- **Privacy Checkup.** Privacy Checkup gives users a prompt to double-check their existing privacy settings and make sure they are still comfortable with them.¹⁷⁴ In January, we updated Privacy Checkup to include:
 - Who can see what you share
 - How to keep your account secure
 - How people can find you on Facebook
 - Your data settings on Facebook.

After the release of the updated Privacy Checkup in January 2020, we sent a prompt to 2 billion Facebook users around the world to encourage them to double check their privacy settings. We provide notifications relating to Privacy Checkup regularly.

- **Off-Facebook Activity.** Off-Facebook Activity, shown in Figure 19, was

¹⁷¹ Meta, *Meta Help Center*, <https://www.facebook.com/help/>; Meta, *Tools to help you control your privacy and security on Facebook*, <https://www.facebook.com/privacy/>

¹⁷² Meta, ‘Introducing Privacy Centre’, *Meta Newsroom*, 7 January 2022, <https://about.fb.com/news/2022/01/introducing-privacy-center/>

¹⁷³ Meta, ‘Introducing manage activity’, *Meta Newsroom*, 2 June 2020, <https://about.fb.com/news/2020/06/introducing-manage-activity/>

¹⁷⁴ Meta, ‘Guiding You Through Your Privacy Choices’, *Meta Newsroom*, 6 January 2020, <https://about.fb.com/news/2020/01/privacy-checkup/>

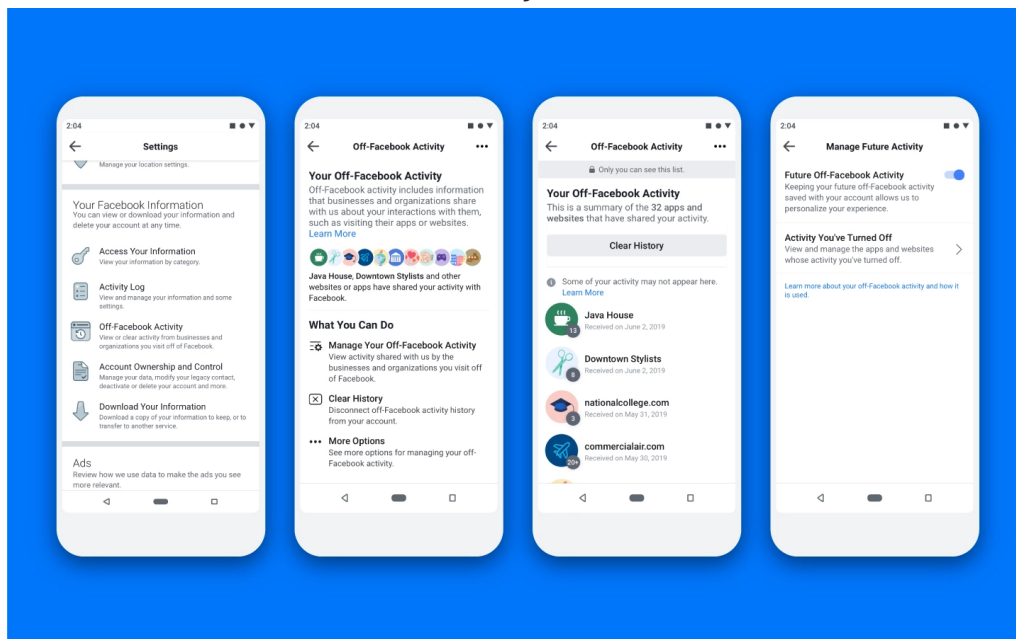
unprecedented when it was launched in 2020, and we believe it remains unmatched today.¹⁷⁵

Some businesses send Facebook information about users' activity on their sites, and we use that information to show ads that are relevant to those users, subject to their privacy settings. Off-Facebook Activity provides users with a summary of that information and gives a control for users to clear that information from their account. Specifically, users can:

- See a summary of the information other apps and websites have sent Facebook through our online business tools, like Facebook Pixel or Facebook Login;
- Disconnect this information from their account; and
- Choose to disconnect future off-Facebook activity from their account. This is possible for all off-Facebook activity, or just for specific apps and websites.

If a user clears their off-Facebook activity, we'll remove the identifying information from the data that apps and websites choose to send us. We won't know which websites they visited or what they did there, and we won't use any of the disconnected data to target ads to that user on our services.

Figure 19: Facebook's 'Off-Facebook Activity' tool



¹⁷⁵ M Zuckerberg, 'Starting the Decade By Giving You More Control Over Your Privacy', *Meta Newsroom*, 28 January 2020, <https://about.fb.com/news/2020/01/data-privacy-day-2020/>

- **Privacy-protective settings and tools for young people.** As explained in the ‘Ensuring Age-Appropriate Experiences’ section above, we offer a number of products, tools and controls that give young people age-appropriate and privacy-protective experiences. These controls put a number of default protections in place for those under the age of 18. They also help to empower young people to make the right choices about their experience online, and the information they want to see and share.

Privacy enhancing technologies

We continue to invest in the research and development of privacy-enhancing technologies (PETs). In August 2021, we announced that we are investing in a multi-year effort, in partnership with academics, global organisations and developers to build new, privacy-enhancing solutions.¹⁷⁶

PETs involve advanced techniques to help minimise the data that’s processed, while preserving critical functionality like ad measurement and personalisation. For example, one PET we have developed, known as On-Device Learning, trains an algorithm from insights processed right on a user’s device without sending individual data to a remote service or cloud. This technology could help us find new ways to show people relevant ads, without needing to ever learn about specific actions individuals take on other apps and websites.

We continue to invest in research and innovation to inform these new PET products. In 2020, we funded USD\$2 million towards research on privacy preserving technologies, user experiences in privacy, and privacy in AR/VR and smart device products. In 2021, we focussed our funding on PETs. Australian researchers have been highly successful in these RFPs, and we continue to work with them to inform our approach to privacy. In 2020, Taeho Jung (University of Notre Dame), Olya Ohrimenko (University of Melbourne) and Kanchana Thilakarathna and Albert Zomaya (University of Sydney) were all granted funding towards their privacy-enhancing research.

¹⁷⁶ Meta, ‘What are privacy-enhancing technologies (PETS) and how will they apply to ads?’, *Meta Newsroom*, 11 August 2021, <https://about.fb.com/news/2021/08/privacy-enhancing-technologies-and-ads/>

Data security

When discussing the security of data, it is important to not only talk about the collection and use of data (as outlined in the Privacy section above), but also to consider the way data is protected throughout its lifecycle.

We take a multi-faceted approach to data security, focussing on areas as diverse as penetration testing, spam prevention, disrupting operations run by adversaries, data protection, and taking legal steps to respond to cyber attacks. We've invested significantly to ensure our network infrastructure is strong, secure and capable of enforcing strong encryption for billions of users.¹⁷⁷ We use a combination of expert teams and automated technology to detect potential abuses of our services.

Below, we provide more detail below on:

- The policies we set for use of our apps and how we enforce those policies.
- Tools we provide to support Australians to protect their data's security.
- Partnerships to encourage collaboration across the data and cyber security environment.

Underpinning our policies, tools and partnerships are our five principles for data security, which aim to ensure we implement the strongest measures to protect users' data.¹⁷⁸

- **Encryption and security.** To protect user's data, we implement a comprehensive security program, including measures such as encryption when data is in transit, to protect user data at all times. We adapt and improve our security to keep ahead of the evolving risks and security threats that we face.
- **No "back door" governmental access.** We do not provide any government with direct access or encryption "back doors." We believe that intentionally weakening our services in this way would undermine the security that is necessary to protect people who use our global service.
- **Robust policies.** For a long time, we have had comprehensive policies in place governing how we evaluate and respond to government requests for user data. We review each request and only provide information in response to requests that we determine are valid, producing only information that is narrowly tailored

¹⁷⁷ N Goyal, A Asogamoorthy and M Yang, 'Enforcing encryption at scale', *Engineering at Meta*, 12 July 2021, <https://engineering.fb.com/2021/07/12/security/enforcing-encryption/>

¹⁷⁸ E Egan, 'Steps we take to transfer data securely', *Meta Newsroom*, 11 March 2021, <https://about.fb.com/news/2021/03/steps-we-take-to-transfer-data-securely/>

to respond to that request.

- **Standing up for our users.** Where government requests are deficient (e.g. overbroad or legally deficient), we push back and engage governments to address any apparent deficiency. We would also challenge any order seeking to require us to redesign our systems in a way that would undermine the security we provide to protect people's data.
- **Providing transparency.** We strive to be open and proactive about the way we safeguard people's privacy, security and access to information online. For this reason, it is our policy to notify users of requests for their information prior to any disclosure, unless we are prohibited by law from doing so, or in exceptional circumstances when notice would be counterproductive such as when a child is at risk of harm.

As mentioned in the 'Transparency and Accountability' section above, we publish biannual transparency reports concerning the nature and extent of government requests we receive for user data.

Policies and enforcement

Our Community Standards prohibit inauthentic accounts or behaviour that intends to mislead users. Specifically, our Community Standards contain requirements about: cybersecurity. We specifically have a policy that users cannot: attempt to compromise user accounts, profiles or other Facebook entities; attempt to gain authorised access; gather sensitive information via deceptive means; or abuse our products and services.

We also use legal recourse against those who violate our policies to perpetrate cyber security risks. In the past year, we've taken over 300 enforcement actions against people who abused our platforms.¹⁷⁹ These actions can include sending cease and desist letters, disabling accounts, filing lawsuits or requesting assistance from hosting providers to have accounts taken down.¹⁸⁰ For example, in December 2021 we filed a federal lawsuit in California to disrupt phishing attacks designed to deceive people into sharing their login credentials on fake login pages for Facebook, Messenger, Instagram and WhatsApp.¹⁸¹

¹⁷⁹ M Clark, 'Scraping by the Numbers', *Meta Newsroom*, 19 May 2021, <https://about.fb.com/news/2021/05/scraping-by-the-numbers/>

¹⁸⁰ Other examples can be found at: <https://about.fb.com/news/2020/06/automation-software-lawsuits/>

¹⁸¹ J Romero, 'Taking legal action against phishing attacks', *Meta Newsroom*, 20 December 2021, <https://about.fb.com/news/2021/12/taking-legal-action-against-phishing-attacks/>

Most recently, we reported on our investigation into seven different surveillance-for-hire entities. The global surveillance-for-hire industry targets people across the internet to exploit vulnerability, collect intelligence, manipulate them into revealing information and compromise their devices and accounts. Following our months-long investigation, we disabled seven entities based in China, Israel, India and North Macedonia who targeted people across the internet in over 100 countries.¹⁸²

There are a number of cyber security threats where we have taken action to protect Australians. One such example is the actions we took in March 2021 against a group of hackers in China known in the security industry as Earth Empusa or Evil Eyewho targeted the Uyghur diaspora in a number of countries, including Australia.¹⁸³ Through a combination of our security and detection measures, we were able to identify a number of cyber espionage tactics and ultimately disrupt their ability to use their infrastructure to abuse our platform, distribute malware and hack people's accounts across the internet. In announcing our detection of this network, we also shared threat indicators - such as malware hashes and malicious domains used - to enable other companies and platforms to detect and stop this activity.

Tools

Users also play an important role in protecting themselves and their data. We make a number of tools available to support users to protect their cyber security. We've also built a dedicated hub for users outlining the steps we take to protect privacy and security and the tools they can use.¹⁸⁴

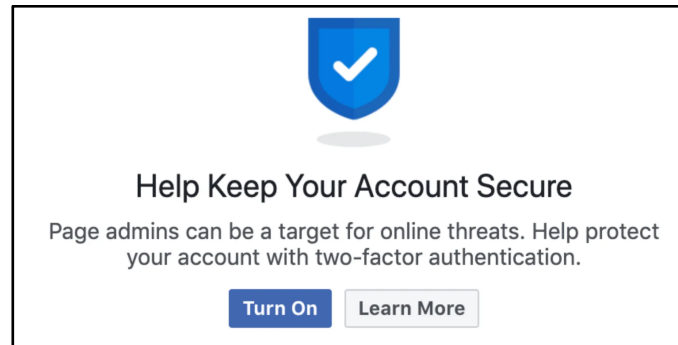
- **Proactive reminders.** We regularly provide in-product reminders to prompt users to strengthen the security of their account, shown in Figure 20 below.

¹⁸² M Dvilyanski and N Gleicher, 'Taking Action Against Hackers in China, *Facebook Newsroom*, 24 March 2021, <https://about.fb.com/news/2021/03/taking-action-against-hackers-in-china/>

¹⁸³ M Dvilyanski and N Gleicher, 'Taking Action Against Hackers in China, *Facebook Newsroom*, 24 March 2021, <https://about.fb.com/news/2021/03/taking-action-against-hackers-in-china/>

¹⁸⁴ See the 'Safe and Secure' Hub, <https://www.facebook.com/about/basics/stay-safe-and-secure>

Figure 20: Facebook security prompt



- **Providing accessible information on how to keep your account secure.** We offer easily accessible security tips for both Facebook¹⁸⁵ and Instagram¹⁸⁶, including an in product step-by-step guide to conduct a Security Checkup on your account.
- **Mandating protections for highly targeted users.** In December 2021, we expanded the Facebook Protect program to Australia.¹⁸⁷ The Facebook Protect program helps groups that are likely to be targeted by malicious hackers - including human rights defenders, journalists, and government officials - to adopt stronger account security protections, like two-factor authentication, and monitors for potential hacking threats. Since rolling out Facebook Protect, we saw a 90 per cent adoption rate in these highly targeted groups in the first month.

Partnerships

We partner with industry, regulators and government to share our findings on threat actors, and raise awareness.

Partnerships to share intelligence

External security researchers are key partners for us. Since 2011, we have encouraged security researchers to responsibly disclose potential issues through our Bug Bounty program.¹⁸⁸ In 2021 alone we awarded over \$2.3 million to researchers from more than 46

¹⁸⁵ Meta, *Security Features and Tips*, <https://www.facebook.com/about/security>

¹⁸⁶ Meta, 'Instagram Security Tips', *Instagram Help Centre*, [https://help.instagram.com/369001149843369/?helpref=hc_fnav&bc\[0\]=Instagram%20Help&bc\[1\]=Privacy%20Safety%20and%20Security&bc\[2\]=Login%20and%20Passwords](https://help.instagram.com/369001149843369/?helpref=hc_fnav&bc[0]=Instagram%20Help&bc[1]=Privacy%20Safety%20and%20Security&bc[2]=Login%20and%20Passwords)

¹⁸⁷ N Gleicher, 'Expanding Facebook Protect to more countries', *Meta Newsroom*, 2 December 2021, <https://about.fb.com/news/2021/12/expanding-facebook-protect-to-more-countries/>

¹⁸⁸ D Gurfinkel, 'Marketing the 10th Anniversary of Our Bug Bounty Program', *Meta Newsroom*, 19 November 2020, <https://about.fb.com/news/2020/11/bug-bounty-program-10th-anniversary/>

countries and have received around 25,000 reports in total, issuing bounties on over 800 of these.

As the threat landscape has evolved, we have continued to develop the Bug Bounty program. In December 2021, we expanded the program to cover new forms of scraping - scraping bugs and scraping databases.¹⁸⁹

We may also occasionally find critical security bugs or vulnerabilities in third-party code or systems when we interact with them. In some instances, there may be significant complexity in working through how to resolve the bug with the partner. We have a Vulnerability Disclosure Policy that sets out how we approach these situations.¹⁹⁰

We have also facilitated industry efforts to combat cyberthreats through threat signal sharing between industry peers through our ThreatExchange API platform, which we launched in 2015.¹⁹¹ This program supports the sharing of threat information (e.g., malicious domains hosting malware, phishing scams, malware hashes) to help security professionals better tackle cyber threats by learning from each other's discoveries.

Partnerships to raise awareness

We raise awareness of cyber security issues through involvement in public campaigns, such as Scam Awareness Week. In 2021, to mark Scam Awareness Week, we ran a dedicated campaign to raise awareness for the most common scams online, and provide tips on how to avoid them.¹⁹² In addition to the campaign, we've supported IDCARE to promote their Cyber Resilience Outreach Centre (CROC) program across our platforms. The CROC program will take mobile cyber clinics to remote and rural locations in Australia, and provide scam and cyber threat and account protection training. The CROC program began in October and will deliver up to 50 clinics across Australia.

¹⁸⁹ D Gurfinkel, 'Expanding our bug bounty program to address scraping', *Meta Newsroom*, 15 December 2021, <https://about.fb.com/news/2021/12/expanding-bug-bounty-program-to-address-scraping/>

¹⁹⁰ Meta, *Vulnerability Disclosure Policy*, <https://www.facebook.com/security/advisories/Vulnerability-Disclosure-Policy>

¹⁹¹ Meta, 'Welcome to ThreatExchange', *Facebook for Developers Help Centre*, <https://developers.facebook.com/docs/threat-exchange/getting-started/>

¹⁹² Meta, 'Staying safe online - scam prevention', *Facebook Australia Blog*, <https://australia.fb.com/staying-safe-online/>

Current and future regulation

Governments around the world are developing and implementing new rules for the internet, across a range of policy areas.

Since 2019, our CEO has been calling for new rules for the internet, especially in areas such as content and online safety, privacy, elections and data portability.¹⁹³ We have backed that up by encouraging the global debate on best practice regulation, for example, by releasing a white paper on content governance in February 2020.¹⁹⁴

Consistent with this global commitment, we have supported and encouraged regulation in Australia. For example: we were the first company to publicly endorse the eSafety Commissioner's Safety by Design Guidelines; we funded expert research on best practice misinformation regulation by an Australian academic in February 2021; and we were a critical driver in landing a world-leading industry code on misinformation and disinformation here in Australia.

This call for new regulations comes in addition to the considerable work that has already been done to respect local laws with respect to content regulation, where local Australian laws and expectations differ to our global policies, and our work to work constructively with local regulators.

The Australian Government has been active in introducing new regulation specifically related to digital platforms. In the last three years, at least **14** new digital platforms regulations have been pursued at the federal level, including the following non-exhaustive list:

- The Online Safety Act (including underpinning regulations such as the Basic Online Safety Expectations, the Restricted Access System Declaration, and associated mandatory industry codes)
- The Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act
- Enhancing Online Safety (Non-consensual Sharing of Intimate Images) Act
- The Social Media (Anti-Trolling) Bill
- Legislation for an Online Privacy Code specific to digital platforms
- Upcoming reform of the Privacy Act 1988
- The News Media and Digital Platforms Mandatory Bargaining Code

¹⁹³ M Zuckerberg, 'The internet needs new rules, let's start in these four areas', *Washington Post*, 30 March 2019, https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html

¹⁹⁴ M Bickert, 'Charting a way forward on online content regulation', *Meta Newsroom*, 17 February 2020, <https://about.fb.com/news/2020/02/online-content-regulation/>

- An industry code on disinformation and misinformation (instigated at the Government's request and delivered with oversight from the Australian Communications and Media Authority)
- The Telecommunications and Other Legislation Amendment (Assistance and Access) Act
- The International Production Orders Act
- Surveillance Legislation (Identify and Disrupt) Act
- Security of Critical Infrastructure Act and associated reform in 2021,

This legislation is on top of existing regulations that cover digital platforms, including online safety, privacy, consumer protection and multinational taxation laws.

In addition, consultation is currently being undertaken with respect to two privacy reform processes that have been proposed to address youth mental health, and an anti-trolling/defamation exposure draft legislation which is also designed to address concerns around harmful online content. Additional regulations have been foreshadowed by the Government, including digital platforms-specific competition laws; age or identity verification; and new obligations about working with law enforcement.

Simultaneously, the Government and Parliament have inquired into digital platforms related issues through at least **18** inquiries over the last three years, including:

- The Digital Platforms Inquiry, undertaken by the Australia Competition and Consumer Commission
- The Digital Platforms Services Inquiry, undertaken by the Australia Competition and Consumer Commission, and still underway
- Senate Select Committee on Foreign Interference Through Social Media
- The Parliamentary Joint Committee on Intelligence and Security's inquiry into extremist moves and radicalisation
- The Parliamentary Joint Committee on Intelligence and Security's post-passage review of the Assistance and Access Legislation
- The Parliamentary Joint Committee on Law Enforcement's inquiry into child exploitation online
- The Parliamentary Joint Committee on Law Enforcement's post-passage review of the Abhorrent Violent Material legislation
- The Senate Environment and Communication Committee's inquiry into media diversity in Australia
- Consultation on strengthening Australia's cyber security regulations and incentives

- and associated Parliamentary Committees and consultation processes for the legislation named above.

Meta has responded constructively to these inquiries, and we have supported many of the new laws which have resulted from them (including, principally, the Online Safety Act).

However, we list these new laws and inquiries to make this point: it is no longer accurate to say that digital platforms are unregulated in Australia or that the risks of the internet are unexamined. The issue for policymakers is no longer a lack of regulation, but whether the existing regulations are effective.

There are also new risks for policymakers to consider. There are two we raise in particular:

1. **Policymakers should be alive to the risk of overlapping, duplicative or inconsistent rules in different laws.** Indeed, many of the online safety-related laws and regulations that have already been passed by Parliament are yet to be implemented. The Online Safety Act - including the Basic Online Safety Expectations and Restricted Access System Declaration - take effect later in January 2022. Industry codes on online safety come into effect in July 2022. There is a massive effort across the industry to both prepare their compliance efforts for these laws and to constructively contribute to the development of online safety codes.
2. **The overall regulatory approach taken by Australia needs to be viewed in the context of a global contest for competing visions of the internet.** Other countries do look to Australia, and it is important to consider whether we are setting an exemplar that encourages a liberal, open and democratic approach to the internet, or an internet that is more closed, tightly controlled and fragmented.

With respect to any new regulatory proposals that the Committee may wish to consider, we encourage the Committee to craft any recommendations mindful of: the existing Australian laws that relate to digital platforms; the potential for regulatory inconsistency given the very considerable digital regulation reform process currently underway; and the current state of internet fragmentation to ensure that any new regulatory proposals are consistent with the liberal democratic values that Australia is advocating across the region.

Current regulatory landscape

Meta has been responsive to a significant and diverse number of Australian laws and regulators, for many years.

Over the last decade, Meta in Australia has responded to questions, requests for further information and complaints from regulators and bodies including the Office of eSafety Commissioner, the Privacy Commissioner, the Australia Competition and Consumer Commission, and state and territory consumer protection agencies, the Australian Electoral Commission and state and territory electoral bodies, the Australian Securities and Investments Commission, the Australian Human Rights Commission, the Australian Tax Office, all state, territory and federal police, and many others. This is in addition to ongoing work with education departments, schools, child safety and mental health organisations and also civil society organisations.

In response to these complaints, we may action content and accounts consistent with our global policies, restrict access to content out of respect for local laws or ask for further clarification about local law concerns. With respect to content restrictions where local law compliance requires a different standard to our global policy enforcement, the transparency data indicates that we have taken action to restrict content out of respect for Australian law since we first started releasing data on our content restrictions in July 2014.¹⁹⁵

These content restrictions are in addition to global policy actions that we have taken in response to complaints from Australian consumers and regulators.

Regulatory inconsistency

Given the technology industry now faces significantly more regulation and there is an active reform agenda currently underway, there is greater risk of overlapping, duplicative or inconsistent rules in different laws.

Take age verification as a case study.

The Government is pursuing multiple different regulations that contain slightly similar but differing objectives. The possibility of age restrictions for social media was introduced in the Online Safety Act, which received Royal Assent in July 2021. As part of that requirement, eSafety released a draft declaration for a Restricted Access System

¹⁹⁵ Meta, *Content Restrictions*, <https://transparency.fb.com/data/content-restrictions/country/AU/>

Declaration with greater detail on age restrictions in August 2021. This is due to take effect in January 2022.

Simultaneously, eSafety began consultation on steps taken to verify the ages of users as part of an Age Verification roadmap in August 2021. This is due to be completed in December 2022.

In September 2021, eSafety released their expectations of the requirements to be contained in industry codes under the online safety legislation. In that paper, they outlined their expectation that there would be mandatory requirements for age-dependent restrictions for certain types of adult sexual content. Given the mandatory nature of the codes, it appears to be a de facto age verification requirement in order to meet these obligations.

In October 2021, the Government began consultation on a set of draft Basic Online Safety Expectations that outlined an expectation companies would take steps to verify the age of their users.

In the same month, draft legislation for an Online Privacy Code was released that went further and included mandatory requirements for social media companies to verify the age of all Australian users. The legislation is expected to be passed in the coming months, with an age verification requirement to take effect six months after that.

The Government has also signaled that it is working towards expectations about identity verification (which would go even further again to verify not just a user's age, but their legal name and identity).

The final result could be more than five separate regulations, all with slightly differing requirements around age restrictions and verification.

Whilst well-intended in terms of the outcomes being pursued, the potential for regulatory uncertainty and inconsistency that not only confuses industry but also consumers is significant. Time and care should be taken to ensure requirements are carefully designed, allow companies sufficient time to build for compliance, and be compatible with existing rules.

Similarly, we are concerned that aspects of the Government's proposal to develop an Online Privacy Code will cause inconsistencies with its broader, economy-wide reform of

the Privacy Act. This is because the industry specific Code is being developed prior to the cross-economy law reform.

The Government has released an exposure draft of legislation requiring the development and approval of an Online Privacy Code within 12 months. The Code would apply to social media platforms, data brokerage services and other large online platforms. While the Code itself is to be developed by industry in the first instance, the enabling legislation contains a number of prescriptive requirements in relation to the Code. These requirements include that the Code must:

- set out detail on how social media platforms are to comply with the existing Privacy Principles, including detail around what constitutes valid consent;
- provide individuals with a right to object to the use or disclosure of their information;
- require social media platforms to take all reasonable steps to verify the age of their users, and collect verified parental consent prior to collecting personal information of users under the age of 16; and
- require the collection, use and disclosure of personal information relating to a person under 18 to be ‘fair and reasonable’.

At the same time, the Government has also released a Privacy Act Review Discussion Paper, which contains more than sixty proposed amendments to the Australian Privacy Act. If the Online Privacy Code is developed first, there are likely to be inconsistencies because:

- The Code will be developed by reference to current Australian Privacy Principles, which may then be amended due to proposals contemplated by the Discussion Paper. This would make the Code immediately out of date;
- The Code would introduce certain requirements in relation to organisations it binds, but these same requirements may be implemented into an amended Privacy Act and apply across the economy. This would make these provisions in the Code redundant at best or, if formulated in slightly different ways, inconsistent with the Privacy Act.

We urge Australian policymakers to ensure any recommendations for future regulation relating to digital platforms fully consider the entire suite of existing and planned regulation in this space.

Combating fragmentation of the internet

In addition to a narrow focus on the interplay of domestic legislation, we encourage the Committee to consider any further regulatory measures against the broader geo-political context and state of the global internet.

The origins of the global internet that has given rise to the tremendous ability for Australians to connect and small Aussie businesses to thrive, was founded on liberal, democratic principles and an open internet, pioneered by US companies – one of Australia’s closest allies. However, the values that underpin the original global internet are increasingly being challenged by a different model of the internet pioneered by other strong forces in the region – a heavily surveilled closed internet, with data localisation, and very little individual privacy.

This is why Meta has been calling for a “Bretton Woods” moment for the internet¹⁹⁶ – the creation of a multilateral, international framework for the internet that would agree some inviolable principles of how the global internet operates – such as privacy of the individual, user rights, open data flows across borders, transparency and accountability by which the systems are operated, strict limits on the amount of sort of intrusive censorship, agreement on whether governments or industries – among other principles that accord with the liberal democratic origins of the global internet.

We encourage the Committee to make sure any proposals for further internet regulatory reform take account of the contest for the future of the global internet and avoid further fragmentation.

Summary of recommendations

1. We encourage the Government to establish a process to consider what ‘success’ looks like in relation to online harmful content. A collaborative industry-led process could assist with developing metrics, to better understand whether regulations and industry efforts have been genuinely effective.
2. The Government should require statutory reviews of new digital platforms legislation after it’s passed to ensure it is effective and fit-for-purpose (eg. a review to commence 18 months after legislation has been passed). Given the significant amount of new legislation that has been passed recently, we suggest

¹⁹⁶ N Clegg, ‘A Bretton Woods for the digital age can save the open internet’, *Australian Financial Review*, 16 November 2021, <https://www.afr.com/technology/a-bretton-woods-for-the-digital-age-can-save-the-open-internet-20211115-p5994h>

the Government could commission a holistic independent stocktake and review of the effectiveness of existing digital platforms regulation (primarily the Online Safety Act) in 2023.

3. For forthcoming regulations that may overlap or duplicate existing requirements (like the Online Privacy Code and age verification requirements), the Government should amend existing draft legislation to reduce the risk of misalignment across multiple regulations.
4. The Australian Government should work through international fora to work towards the establishment of Bretton Woods-style infrastructure for the digital age.
5. We encourage the Committee to make sure any domestic proposals for further internet regulatory reform take account of the contest for the future of the global internet. Australia should work to set an exemplar that encourages a liberal, open and democratic approach to the internet.