Meta

Submission to the *Statutory Review of the Online Safety Act 2021*

JULY 2024

Executive Summary

Meta welcomes the opportunity to provide a submission to the Statutory Review of the Online Safety Act 2021 (**the Statutory Review**).

<u>Australia's online safety framework</u>

Whilst Australia's online safety framework started from 2015 as a complaint-based, notice and takedown scheme, in the years since it has been added to on an ad hoc basis, to expand the scope of content to which it applies. In 2021, the framework was comprehensively reformed to adopt a more hybrid approach that includes a notice and takedown scheme under the primary legislation, principles-based regulation with the *Online Safety (Basic Online Safety Expectations) Determination 2022* (**BOSE**), and systems-based regulation under industry codes and standards. Given the ad hoc approach to updating Australia's online safety regulations, the Statutory Review is timely and can help ensure that the *Online Safety Act 2021* (Cth) (**OSA**) remains fit for purpose in holding industry to account and ensuring that all Australians, including young Australians, have a safe and positive experience online.

In sum, Meta supports the hybrid approach adopted for Australia's online safety framework subject to our comments below in this submission, and further encourages the Statutory Review to make recommendations that improve, rather than substantially adjust, this approach. We also encourage the Statutory Review to take account of the existing investments within industry when framing its recommendations to ensure that Australia's online safety framework leverages existing investments and drives consistency across industry.

The importance of consistent expectation setting and enforcement across industry is paramount in the modern internet age because malicious actors will work to spread across many internet services in an attempt to ensure that a takedown by any one of these services will not disrupt all of their online activities.

Within Meta, since the Australian Parliament first introduced online safety legislation in 2015, the biggest change in Meta's investment in safety and security has been in the increased use of automation and artificial intelligence (**AI**) in our content governance and integrity systems. Increasingly, we have been deploying proactive detection technology to identify and action harmful content before anyone reports it to us. For many categories, our proactive rate (the percentage of content we took action on that we found before a user reported it to us), is more than 99 per cent across high-risk content types.

It is therefore no longer accurate to state that the burden of online safety regulation falls on the user. The correct focus for the Statutory Review should instead be on how to create a regulatory framework that: incentivises investment in safety and security and not just compliance; respects the multi-stakeholder nature of online safety responsibility; and, provides the Office of the eSafety Commissioner with the tools necessary to seek enforcement and accountability across all of industry.

Meta's investments in online safety

At Meta, we are committed to protecting the safety of people when they use Meta's services, and especially the safety of young people. It is essential to our business: Australians and other people around the world will only continue to use our platforms if they feel welcome and safe.

Effective online safety schemes require collaboration between industry, government and the community, who all have a role to play. To uphold our responsibility, we make significant investments in our ability to keep people safe. This includes investing in ongoing policy development, automated and human enforcement of our policies, awareness and educational initiatives, partnerships, as well as tools that allow people to customise their experience on our services, over and above the baseline safety and security efforts we deploy. We've invested more than US$20 billion (~AU$30 billion) on safety and security since 2016, and this includes US$5 billion (~AU$7.5 billion) in the last year alone.

Recognising the multi-stakeholder roles in promoting online safety, Meta has long been calling for new regulation globally, that is harmonised and takes an ecosystem approach, especially in areas such as content and online safety, privacy, elections and data portability.[1] We recognise that private companies should not be making decisions on important issues alone, and that greater accountability and transparency by digital platforms can give governments and the communities they serve greater confidence about the investments and commitments of companies like Meta, particularly when it comes to online safety.

In Australia, we have worked to implement this global approach by complying with the local regulatory frameworks and guidance. We were the first company to publicly endorse the eSafety Commissioner's Safety by Design Guidelines;[2] we are an active industry participant in the development of the first phase of industry codes under the OSA focused on class 1A and class 1B material (**Phase 1 Codes**); we have responded to multiple rounds of reporting notices under the BOSE; and we have worked constructively with the Office of the eSafety Commissioner to respond to complaints, share product briefings and engage in programmatic activities since its establishment in 2015.

[1] 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

[2] Safety by Design Youth Jam, *Facebook,* August 2019, https://www.facebook.com/MetaAustralia/videos/910843179301219

Key factors in reviewing the OSA

Drawing on our experience of engaging on Australian regulatory frameworks over several years, we respectfully suggest that the Statutory Review considers the following key factors as part of its review:

- **Efficacy:** how to measure the efficacy of Australia's online safety regulatory framework as part of identifying what new reforms may be necessary;
- **Benchmarking:** how to best benchmark the existing regulatory obligations across industry to identify where the regulatory framework can drive improvements;
- **Targeted & proportionate:** which obligations are appropriate and proportionate for which layer of service or device within the digital ecosystem;
- **Flexibility:** as technology changes and people's use of it changes, how the OSA regulatory framework can stay robust and adaptive;
- **Complementarity:** given the significant volume of safety and safety-adjacent reforms in Australia, how proposed OSA reforms will work within the broader digital platform regulatory framework, especially privacy laws in relation to age verification proposals; and
- **Global implications:** how industry standard setting in Australia works within international legal frameworks.

We believe incorporating these factors within any proposed reforms will assist the Statutory Review in developing robust reform proposals that stand the test of time and avoid the ongoing amendments of Australia's online safety framework that have taken place since 2015. This will allow: industry to invest in appropriate safety-driven compliance approaches; government to develop capacity to hold industry to account amidst ongoing technological change; and greater community awareness to be built up about what to expect with respect to online services and how these expectations are enforced.

While there have been several international models developed since the OSA's initial enactment some nine years ago, we should be circumspect about the wholesale importation of overseas models, as models that work in one region may not necessarily work in others and many are still too new to assess their efficacy, and it is important that the Statutory Review take into account the existing Australian landscape and context and the multi-part implementation of Australia's OSA that has been taking place over the past several years.

Age assurance solutions

Within the current OSA framework, we encourage the Statutory Review to consider how Australia can craft a leading global approach with respect to age assurance solutions. Debates around ensuring age-appropriate experiences online are occurring globally but the most effective solutions are not yet agreed on.

Within the digital industry there has been long-standing recognition of the importance of ensuring that young people and adults should have different experiences online. This is why Meta has continued to build on our age assurance tools and technologies over the years. As the technology develops, it is becoming clearer globally and in Australia that we need privacy-protective solutions that will allow industry to provide age-appropriate experiences, that will support parents to be more involved in the services used by young people. Age verification and parental oversight are a cornerstone of developing youth safety legislation around the world, but there are no globally accepted industry standards that dictate how companies should approach age assurance online, nor do current laws provide clear guidance on how industry can approach this in a compliant, privacy-preserving way.

Meta supports enhanced assurance of age for young people, through ecosystem approaches and industry-wide solutions where all apps are held to the same consistent standards, especially to avoid young people flocking to services that are far less safe than those that have invested in age-appropriate protections and experiences. As has already been proposed by the eSafety Commissioner in the ongoing discussion of the second phase of industry codes focused on class 1C and class 2 material (**Phase 2 Codes**), we advocate for an ecosystem approach to age-assurance, with responsibilities and obligations taken by all involved from the operating system and device level down to individual apps and services. We believe that any additional requirements around age assurance should consider the broader ecosystem and be required at the operating system and device level, where age collection and confirmation data can be processed once and thereafter shared across the industry throughout the respective app ecosystems whenever an app is downloaded.

Measures at the operating system and device level will be complementary to, and should not replace, the age assurance efforts and responsibilities that services like Meta are already pursuing. This will yield ecosystem wide benefits as new entrants to app ecosystems ensure age appropriate experiences for all. On this basis, apps and services can take additional measures to directly verify age when they have signals a user may be misrepresenting their age, and also deploy additional ongoing measures to bolster age assurance. By crafting the broader OSA framework cognisant of this opportunity, the Statutory Review can set a global best practice around age-assurance, with the expertise of the eSafety Commissioner.

The wider context

And finally, we encourage the Statutory Review to undertake any reform of the OSA framework mindful of the significant work that companies such as Meta undertake with respect to online safety, above and beyond regulatory requirements, and structure any recommendations in such a way as to leverage these investments. At Meta, we take a layered and expansive approach to safety. Regulatory compliance is only one part of our responsibility to promote online safety for all Australians, especially young Australians. This is why – in addition to our investment in policies, enforcement, and tools – we also work with local online safety and mental health organisations. We have an Australian Online Safety Advisory Group comprising experts such as CyberSafety Solutions, PROJECT ROCKIT, WESNET, ReachOut, and many others, with whom we regularly engage on online safety issues, particularly those relating to young people. Meta also invests in long-standing partnerships with both Australian law enforcement and non-profit organisations to promote greater awareness and understanding of our policies, tools and tips and strategies for staying safe online. In the past 12 months, we have undertaken the initiatives with a range of local partners, including the Butterfly Foundation, PROJECT ROCKIT, Kids Helpline and ReachOut.

As the above demonstrates, Meta will continue to be a constructive partner for Australian policymakers in advancing online safety, and we welcome the opportunity to continue to engage with this Statutory Review.

Table of Contents

Overview of Meta’s approach to safety

Before turning to our specific comments on the Issues Paper, we wanted to first outline – at a high level – our existing approach to and investment in online safety, to provide background to our comments on the specific issues raised. We have also provided more details about our investments on specific areas of online safety also being considered by the Statutory Review in the Appendix.

By way of background, Meta’s investment in safety is considerable and ranges across the development and enforcement of policies, investments in partnerships to promote educational initiatives and receive feedback on issues and trends for further action, as well as tools and technology to adjust our scaled policy enforcement for specific issues or experiences of the people, especially young people, who use our services that require additional support.

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

Our 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

To assist the Statutory Review in understanding the nature of Meta’s investment as it works to finalise its recommendations, we outline – at a high level – our approach to online safety, above and beyond regulatory requirements. However, recognising that many parts of the Issues Paper consider specific issues such as women’s safety, public figures, young people, and mental health and wellbeing – we have included separate sections in the Appendix and cross-referenced these in the body of the submission to provide a more detailed insight into the customised solutions we have adopted to address specific issue areas.

Policies

Our policies, known as our Facebook Community Standards and Instagram Community Guidelines,[3] outline what is and is not allowed on Facebook and Instagram. These policies are informed by a range of values to help combat abuse, including safety as a core value, alongside privacy, authenticity, voice, and dignity.[4]

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

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

Meta's policies are also regularly updated to keep pace with changes happening online and offline around the world. We host a regular Policy Forum meeting to discuss potential changes to our policies and their enforcement. A variety of internal and external subject matter experts participate in this meeting and hear input from external groups. In keeping with our commitment to greater transparency, the minutes of these meetings are made publicly available.[5] A change log of changes made to each policy area is available within the Community Standards.[6]

Enforcement

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

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

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

[3] See Meta, 'Facebook Community Standards', *Transparency Center*, https://www.facebook.com/communitystandards; Meta, 'Instagram Community Guidelines', *Help Center*, https://help.instagram.com/477434105621119

[4] Meta, 'Updating the values that inform our community standards', *Newsroom,* 12 September 2019, https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards

[5] Meta, 'Policy Forum Minutes', https://transparency.meta.com/en-gb/policies/improving/policy-forum-minutes/

[6] See, for example, Meta, 'Facebook Community Standards - Child sexual exploitation, abuse and nudity', *Transparency Center*, https://transparency.meta.com/en-gb/policies/community-standards/child-sexual-exploitation-abuse-nudity/

To that end, we have scaled our enforcement to review millions of pieces of content across the world every day, and use our technology to help detect and prioritise content that needs review. We continue to build technologies like RIO,[7] WPIE[8] and XLM-R[9] that can help us identify harmful content faster, across languages and content type (i.e. text, image, etc.). These technologies alongside our continued focus on AI technologies help us to scale our efforts quickly in keeping our platforms safe.

As part of our ongoing commitment to transparency and accountability, we provide data about our enforcement work in our Community Standards Enforcement Report, which we publish quarterly.[10] This report includes metrics such as how much content we are actioning, and what percentage was detected proactively. Currently, we report these metrics against 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 policies. For many categories, our proactive rate (the percentage of content we took action on that we found before a user reported it to us), is well over 90 percent across high-risk content types such as child exploitation material and terrorist content, as well as violent and graphic content. For example, between January and March 2024, we proactively found and actioned 10.6 million pieces of violent and graphic content on Facebook and 12.1 million on Instagram, 98.7% and 99.4% of which (respectively) we proactively found and actioned before people reported it.

Tools and products

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.

[7] 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/

[8] 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/

[9] 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/

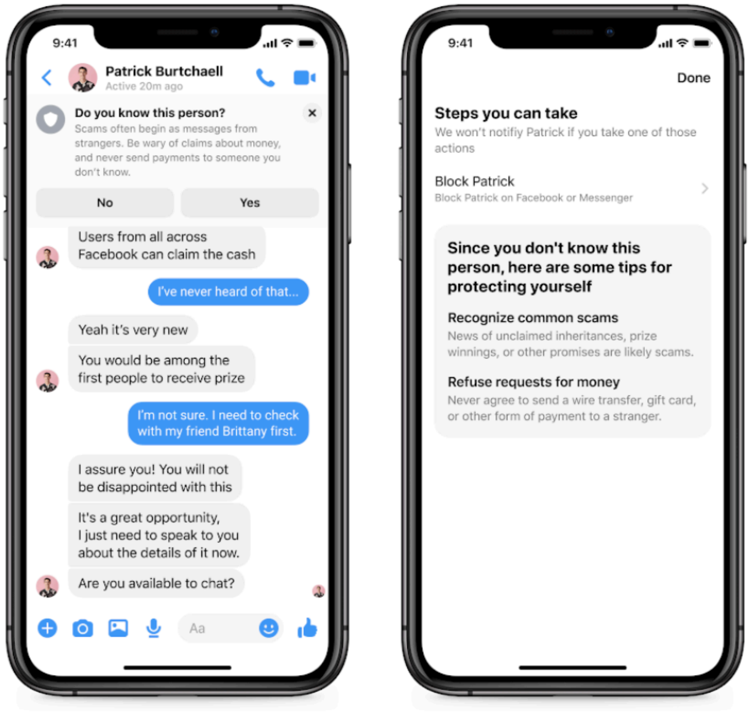
[10] Meta, 'Community Standards Enforcement Report', *Transparency Center*, https://transparency.fb.com/data/community-standards-enforcement/

In addition to the long-standing tools of Block, Report, Hide, Unfollow,[11] 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.[12] 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 have introduced warnings and safety notices across our platforms to educate people on who they are engaging with. For example, in Messenger and Instagram Direct 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.[13] These notices are designed to discourage inappropriate or risky interactions and to limit the potential for harms to occur via Messenger and Instagram.[14]

Figure 1: Messenger Safety Notice



Do you know this person?

Scams often begin as messages from strangers. Be wary of claims about money, and never send payments to someone you don't know.

Users from all across Facebook can claim the cash

I've never heard of that...

Yeah it's very new

You would be among the first people to receive prize

I'm not sure. I need to check with my friend Brittany first.

I assure you! You will not be disappointed with this

It's a great opportunity, I just need to speak to you about the details of it now.

Are you available to chat?

Steps you can take

We won't notifiy Patrick if you take one of those actions

Block Patrick

Block Patrick on Facebook or Messenger

Since you don't know this person, here are some tips for protecting yourself

Recognize common scams

News of unclaimed inheritances, prize winnings, or other promises are likely scams.

Refuse requests for money

Never agree to send a wire transfer, gift card, or other form of payment to a stranger.

[11] An overview of these and other tools is available in the Facebook Safety Center: https://www.facebook.com/safety/tools

[12] Meta, Raising the standard for protecting teens and supporting parents online, *Newsroom*, 7 December 2021, https://about.instagram.com/blog/announcements/raising-the-standard-for-protecting-teens-and-supporting-parents-online

[13] Meta, 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

[14] Meta, '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

- **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[15], and these settings are defaulted to stricter and more restrictive settings for younger users. 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 prohibit content that promotes or encourages self-harm and eating disorders, we do allow people to share their own experiences 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, we make potentially sensitive content harder to find - by restricting search results, removing related content from recommendation surfaces, and applying sensitivity screens to individual pieces of content, while pointing people to helpful resources. This includes directing people to dedicated resources, including, in Australia, from the Butterfly Foundation.[16]

Resources

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 Hub[17] and the Facebook Safety Center.[18] These also include the:

- Bullying Prevention Hub developed in partnership with the Yale Centre for Emotional Intelligence;[19]

- 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;[20]

[15] Meta, 'Our commitment to keeping people safe', *Newsroom,* 11 February 2020, https://about.instagram.com/blog/announcements/making-instagram-safer-for-the-youngest-members-of-our-community
[16] Instagram, 'How we're supporting people affected by eating disorders and negative body image', *Newsroom*, 22 February 2021, https://about.fb.com/news/2021/02/supporting-people-affected-by-eating-disorders-and-negative-body-image
[17] Instagram, *Instagram Safety*, https://about.instagram.com/safety
[18] Meta, *Digital Literacy Library*, https://www.facebook.com/safety/educators
[19] Meta, *Bullying Prevention Hub*, https://about.meta.com/actions/safety/topics/bullying-harassment
[20] Meta, *Youth Portal,* https://www.facebook.com/safety/youth?locale=en_GB

- Family Centre, which provides a central place for families on the various tools and third-party educational resources that support parental and guardian involvement in their teens' social media usage,[21]
- 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;[22]
- Suicide Prevention Support Centre that provides resources and guidance on how to access and offer support.[23]

Partnerships

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

Globally, we collaborate across industry through organisations like the Tech Coalition, an industry association dedicated solely to eradicating child sexual exploitation and abuse online.[24] In November 2023, we supported the Tech Coalition to establish the Lantern program,[25] which enables technology companies to share signals about accounts and behaviors that violate their child safety policies. As a founding member of Lantern, we provided the Tech Coalition with the technical infrastructure that sits behind the program as well as oversee the technology with them, ensuring it is simple to use and provides our partners with the information they need to track down potential predators on their own platforms. This builds on the work of Project Protect[26] – an industry effort launched in 2020 to combat online child sexual abuse.

We are also a part of the WePROTECT Global Alliance[27] industry committee. WeProtect brings together experts from government, the private sector and civil society to protect children from sexual exploitation and abuse online.

We have also convened a Global Safety Advisory Council[28], which comprises leading safety organisations and experts from around the world. Council members provide expertise and perspective that inform Meta's approach to safety. The Australian youth anti-bullying organisation PROJECT ROCKIT is one of 11 organisations globally that serves on this Council.

[21] Meta*, Family Center,* https://familycenter.meta.com/

[22] Meta, *Get Digital Hub*, https://www.facebook.com/fbgetdigital

[23] Meta, *Suicide Prevention Support Centre,* https://www.facebook.com/safety/wellbeing/suicideprevention

[24] Tech Coalition, https://www.technologycoalition.org

[25] Meta, 'Introducing Lantern: Protecting Children Online', *Newsroom*, 7 November 2023, https://about.fb.com/news/2023/11/lantern-program-protecting-children-online

[26] Meta, 'Facebook Joins Industry Effort to Fight Child Exploitation Online'. *Newsroom*, 11 June 2020, https://about.fb.com/news/2020/06/fighting-child-exploitation-online

[27] WeProtect Global Alliance, https://www.weprotect.org

[28] Meta, 'Learn more about the Meta Safety Advisory Council', *Help Center*, https://www.facebook.com/help/222332597793306

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.[29] Digital Ambassadors is a youth-led, peer-based anti-bullying initiative. A Digital Ambassador aims to utilise strategies to safely connect and tackle online hate. This is a more than decade-long- partnership that has directly empowered more than 25,000 young Australians to tackle cyberbullying.[30]

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

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

- **Butterfly Foundation:** In May 2024, we launched ‘Enter the Chat’, an education campaign that brought together a group of Australian creators to discuss the impact that certain types of online content may have on body image, how to create content more consciously and what safety tools are available on Instagram to support body image and wellbeing.[31]

- **PROJECT ROCKIT**: In November 2023, we partnered with youth-driven organisation PROJECT ROCKIT to create ‘Intimate Images Unwrapped’, a series of educational videos that aimed to build greater literacy and awareness around the dynamics of sharing of intimate images.[32]

- **Kids Helpline and ACCCE:** In November 2023, we partnered with the Australian Federal Police-led Australian Centre to Counter Child Exploitation, Kids Helpline and US-based organisation NoFiltr (Thorn) to inform young people about sextortion. The campaign included educational resources encouraging preventative behaviours online, the signs to look out for, where to report and where to seek support.[33]

- **ReachOut:** In 2023, we partnered with youth mental health service, ReachOut, to launch a creator-led campaign aimed at fostering social and emotional wellbeing in the lead-up to, and following, the Voice to Parliament referendum. The campaign focused on supporting and empowering young First Nations people in navigating the complex social

[29] PROJECT ROCKIT, *Launching: Digital Ambassadors,* https://www.projectrockit.com.au/digitalambassadors
[30] R Thomas, ‘Young People at the Centre’, Meta Australia Policy Blog, *Medium*, 8 February 2021 (updated 28 January 2023), https://medium.com/meta-australia-policy-blog/young-people-at-the-centre-25142d16c0cf
[31] Butterfly Foundation, *Enter the Chat*, https://butterfly.org.au/get-involved/campaigns/enterthechat
[32] PROJECT ROCKIT, *Intimate Images Unwrapped*, https://www.projectrockit.com.au/intimate-images
[33] NoFiltr, https://nofiltr.org

and emotional wellbeing challenges resulting from the referendum and its surrounding debate.[34]

In addition to this general overview of our approach to safety, in the Appendix we outline our approach to safety in relation to particular groups within our community and issues in respect of which there has been recent public debate and that have been flagged for special mention the Statutory Review – young people and parents, and relatedly ensuring age appropriate experiences online, women's safety, public figures, as well as mental health and wellbeing.

[34] ReachOut, *'ReachOut collaborates with First Nations Creators to amplify Social and Emotional Wellbeing support', 6 October 2023,* https://about.au.reachout.com/blog/reachout-collaborates-with-first-nations-creators-to-amplify-social-and-emotional-wellbeing-support

Comments on Australia’s online safety regulatory framework

Australia’s online safety regulatory approach

Meta recognises that digital platform regulation should be premised on platform accountability and responsibility, as well as consumer empowerment and transparency. But set against this is the need to take a harmonised, ecosystem wide approach towards addressing the harms that may arise online.

Given the many new laws that have been enacted or proposed in Australia that relate specifically to digital platforms, the focus now should be on what will be most effective at driving appropriate industry-wide investment in safety and security. Based on our experience of engaging in the Australian regulatory processes over many years, we respectfully suggest that the Statutory Review considers the following key factors as part of its review:

- **Efficacy:** how to measure the efficacy of Australia’s online safety regulatory framework as part of identifying what new reforms may be necessary;
- **Benchmarking:** how to best benchmark the existing regulatory obligations across industry to identify where the regulatory framework can drive improvements;
- **Targeted & proportionate:** which obligations are appropriate and proportionate for which layer of service or device within the digital ecosystem;
- **Flexibility:** as technology changes and people’s use of it changes, how the OSA regulatory framework can stay robust and adaptive;
- **Complementarity:** given the significant volume of safety and safety-adjacent reforms in Australia, how proposed OSA reforms will work within the broader digital platform regulatory framework, especially privacy laws in relation to age verification proposals; and
- **Global implications:** how industry standard setting in Australia works within international legal frameworks.

We believe incorporating these factors within any proposed reforms will assist the Statutory Review in developing robust reform proposals that stand the test of time and avoid the ongoing amendments of Australia's online safety framework that have taken place and continued since 2015. This will allow industry to invest in appropriate safety-driven compliance approaches, government to develop capacity to hold industry to account amidst ongoing technological change, and greater community awareness to be built up about what to expect with respect to online services and how these expectations are enforced.

We thus respectfully offer specific comments relating to the operation and effectiveness of the OSA regime to date, as well as the various issues mentioned in the Statutory Review's Issues Paper.

Ongoing flux in the online safety regulatory landscape

Several of the questions (specifically those considered in ***Part 2, Questions 1 – 7***) considered by the Statutory Review focus on whether the scope, focus or target level of the existing OSA framework is appropriate. For example, the Statutory Review asks whether the objects, the BOSE, and the industry codes should be adjusted or strengthened. However, consistent with our suggestion that any new amendments are considered against the principle of efficacy, we suggest that it is too premature to consider further amendments while many recent amendments are still being implemented, or reforms only recently introduced.

For example, the Phase 1 Codes were only drafted and registered in the past 12 months, and the two industry standards (**Phase 1 Standards**) were only registered this past month in June 2024. All of these contain annual reporting requirements and various risk assessments that have not yet been completed. In addition, the development of the Phase 2 Codes has only just commenced in July 2024 with an indicative target date of drafting to be completed in December 2024.

Given the Phase 1 Codes and Phase 1 Standards have only just taken effect, and the drafting of the Phase 2 Codes is now underway, it seems too early to consider how these processes could be improved. The codes represent significant systems-based safety regulations that have brought together numerous companies represented by five to six industry associations, to agree on standard setting for services within eight sectors of industry. Consequently, we suggest that the focus of the Statutory Review be on identifying measurement frameworks to provide a basis upon which to consider the efficacy of these components at a future date, after they have been allowed to take effect.

Similarly, it is challenging to identify ways in which the BOSE can or should be strengthened given it is currently being implemented, and when newly enacted reforms have only just taken effect and their practical impact can not yet be measured. The BOSE was amended last month, in June 2024, by the *Online Safety (Basic Online Safety Expectations) Amendment Determination 2024* (Cth) (**BOSE Amendments**).[35] This comes after less than two years of operation. During this short timeframe, in September 2022, Meta was one of several companies that responded to the first round of transparency notices, and we have been in the process of responding to a second round of notices since March 2024.

[35] *Online Safety (Basic Online Safety Expectations) Amendment Determination 2024* (Cth)

The BOSE is designed to be a principles-based regime to guide and shape safer outcomes across the industry, allowing different services to adapt their compliance in ways that are positive for their users and the community. As a principles-based regime, the BOSE serves as an anchor for the industry codes and standards, which build on the BOSE principles and expectations and articulate these into specific outcomes and requirements. This creates clarity in objective and intent for industry to build for compliance and is compatible with the structure adopted by other online safety regulations, such as the new UK Online Safety Act, which are premised on principles that are expanded in codes.

However, the purpose and value of the BOSE - designed to be a durable and technology-neutral standard for industry - risks being undermined by further amendments. This shifts the BOSE from this anchoring, principles-based function and makes it more challenging for industry to adjust to changing goal posts (noting some systems, such as reporting systems, take time to build and implement). Introducing granular and technology-specific changes via amendments in the BOSE undermines this BOSE principles-based approach. We support maintaining the original intent of BOSE as a principles-based regime that establishes the norms for safety expectations and gives industry clear north-star safety objectives to continually invest in over time; we respectfully discourage further repeated amendments that would make compliance challenging because of the lack of certainty.

There has also been movement on safety-adjacent laws and regulatory frameworks such as the review and reform of the National Classification Scheme, the Government's age assurance technology pilot, and the forthcoming privacy reforms (which have previously contemplated the development of a 'Children's Online Privacy Code Code' governing privacy for children's data).

Given many of the laws and associated amendments are relatively new, we recommend that the Statutory Review consider any further reforms against the backdrop of all that is occurring, and how improvements can be introduced especially to sequence, complement, or ideally streamline all the ongoing efforts across the safety regulation landscape. With the ongoing state of reform of Australia's online safety laws, there is a risk of incentivising industry to over-index towards regulatory processes that focus on compliance systems and reporting measures for their own sake. This could occur at the expense of other measures that deliver real and sustained impacts to safety and security, such as adapting approaches quickly to manage for new harm types as they evolve.

To take one example, amendments to the BOSE[36] were introduced in June this year. This was despite the fact that while three rounds of non-periodic reporting notices had been issued, notices had not been issued to all service providers covered by the BOSE. Similarly, there has not been time for notices addressing new areas introduced by the BOSE Amendment to be issued and the responses assessed. Many of these areas are now also being considered within

[36] *Online Safety (Basic Online Safety Expectations) Amendment Determination 2024* (Cth)

the Statutory Review and are covered by the codes, and yet, there has not been sufficient time to identify if there is any evidence of a gap in the BOSE that needed to be strengthened.

This is why we have suggested that the Statutory Review consider factors such as efficacy, benchmarking, and targeted and proportionate measures as part of its review. The overall OSA framework will benefit from evidence-based assessment of efficacy and impact once *all* of industry's response to the more recently introduced BOSE Amendments, Phase 1 Codes and Standards and (still to be developed) Phase 2 Codes have had time to mature and settle.

Operation and effectiveness of the OSA

In ***Part 3, the Issues Paper*** poses many questions directed at assessing the effectiveness of the OSA's notice and takedown complaint handling scheme. Given the ad hoc nature with which new complaint schemes have been introduced over the years, we can see benefit in these being streamlined and made consistent. In addition, we can see benefit from developing definitions of harm types specific to the OSA within the OSA framework and not, as currently occurs, with the National Classification Scheme (which is itself undergoing review).

More broadly, however, Meta supports the hybrid approach adopted for Australia's online safety framework that includes a complaint-based notice and takedown scheme together with systems-based regulations as set out in the industry codes and standards.

Notice and takedown regimes provide a useful role, especially for specific types of content or harms like non-consensual intimate imagery and adult cyberbullying, in complementing systems-based regulatory frameworks. Takedown notices are complementary, given the best and most sophisticated systems in the world can not detect and proactively enforce in 100% of cases due to the highly individual nature of some content (such as cyberbullying). This is also consistent with human rights and considerations of freedom of expression, because the regulator can assess and use its powers against the overall backdrop of considerations, together with checks and balances as appropriate. Additionally, effective use of end-user notices can assist the regulator in promoting awareness of the complaint handling scheme and encourage behavioural change so that online content removal is not the only solution to harmful behaviours online.

At this early stage, we observe challenges in operating the systems-based industry codes and standards regimes of the OSA, which are underpinned by definitions used in the National Classification Scheme. The National Classification Scheme is designed to address material in the broadcast sector that can be reviewed and appropriately classified before it is viewed by the public, which does not translate well to the online content landscape. In particular, the definitions, which include a subjective and detailed assessment of impact and context, are challenging to apply to user-generated content where the scale of review requires clear targets for hybrid moderation, utilising both human review and AI. Meta recommends that the Statutory

Review consider decoupling the OSA from the National Classification Scheme (and its proposed reform). As a better alternative, we propose that clear, standalone definitions be included within the OSA, which are developed specifically for the online context. This approach would provide greater certainty and clarity for industry, users and the regulator.

Role of industry

There are several questions in the Issues Paper (specifically ***Questions 5-7 and 25***) that question the role of industry, such as the extent to which safety should be managed by the terms of use of service providers, who should be responsible for drafting industry codes, and the efficacy of industry dispute resolution processes.

Many of the statements made on these topics within the Issues Paper are not wholly consistent with the actual day-to-day operation of service providers such as Meta, nor made with the full context of the issue in question. We encourage the Statutory Review to take a holistic view of the multiple stakeholders required to contribute to a safe online ecosystem and propose effective benchmarking across industry to identify where regulatory frameworks can drive improvements. As outlined above, how best to benchmark existing regulatory obligations across industry is the appropriate basis on which to identify where the regulatory framework can drive improvements.

Specifically, we note:

- It is not accurate to suggest that online safety is managed solely through a service provider's terms of use (***Question 6*** in the Issues Paper). For many companies, terms and policies are one part of a multi-part framework to promote online safety (as we have explained in detailing our approach to safety) and additional measures include enforcement, tools and products, education resources, and partnerships. We outline these above within our overview of Meta's approach to safety as well as below in the Appendix. Additionally, where there is a difference between policies and Australian law, companies such as Meta will restrict access to such content out of respect for Australian law.[37]

- With the Phase 1 Codes and Standards having only just been finalised and the development of the Phase 2 Codes having only just commenced, the Issues Paper seems premature to consider changes at this stage (***Question 5***). In any event, the scope and complexity of the code process – as well as the ability for the eSafety Commissioner to reject them and develop standards in their place – suggests that there are appropriate failsafes to ensure that industry's work reflects Australian community standards.

[37] See, for example, Meta, 'Content Restrictions Based on Local Law', *Transparency Center*, https://transparency.meta.com/reports/content-restrictions/

- The Issues Paper states that internal dispute resolution or complaint handling processes can be lacking (***Page 42, Question 24***) but says this without citing any evidence and without consideration for both the extensive complaints and appeals handling systems of a company such as Meta's nor the existing work being undertaken by industry to develop codes in response to specific requests by the Australian Government in this regard. Specifically:

  - At Meta, there exists user reporting flows for every piece of content across Facebook[38] and Instagram,[39] with user transparency on enforcement actions taken on any such content, together with the opportunity for users to appeal decisions both to the company and to the Oversight Board (on which we provide more detail below).[40]

  - We maintain a Transparency Center[41] with details of our policies, enforcement, and integrity insights, including a quarterly Community Standards Enforcement Report[42] that provides data on how much harmful content we action, the prevalence of harmful content, our proactive detection rates against content that violates our rules, as well as appealed and restored content.

  - To encourage accountability and oversight of our content decisions, we have established an Oversight Board to make binding rulings about content on our services. This Oversight Board comprises 22 experts in human rights and technology, including the Australian academic Professor Nicholas Suzor from Queensland University of Technology.[43] The Board is entirely independent and hears appeals on Meta's decisions relating to content. We have agreed that the Board's decisions will be binding, and the Board is also able to make recommendations about Meta's policies.[44] The Oversight Board also publishes regular Transparency Reports[45] which provide new details on the Oversight Board's cases, decisions and recommendations. Meta regularly publishes our

[38] See, for example, Facebook, 'Reporting abuse', *Help Center*, https://www.facebook.com/help/1753719584844061/?helpref=hc_fnav

[39] See, for example, Instagram, 'How to Report Things', *Help Center*, https://help.instagram.com/2922067214679225

[40] For more details, see, for example , 'Taking down violating content', *Transparency Center*, https://transparency.meta.com/en-gb/enforcement/taking-action/taking-down-violating-content

[41] See Meta, *Transparency Center*, https://transparency.meta.com/en-gb/

[42] Meta, 'Community Standards Enforcement Report', *Transparency Center*, https://transparency.meta.com/reports/community-standards-enforcement

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

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

[45] See Oversight Board, '*Assessing impact in our transparency reports*', https://www.oversightboard.com/transparency-reports/

responses to the Oversight Board's decisions and on our progress in responding to them.[46]

- Industry (including Meta) is currently working with DIGI to develop a voluntary internal dispute resolution code by July 2024, to improve our transparency and accountability to consumers and small businesses regarding our processes.

- The Issues Paper cites the fifth interim report of the Australian Competition and Consumer Commission (ACCC) Digital Platform Services Inquiry 2020–25 (DPSI No 5 Report) to support recommendations in this area, which were supported in principle by the Government on 8 December 2023.[47] It states that the DPSI No 5 report noted potential for an ombuds scheme [linked to internal dispute resolution obligations] to cover broader online disputes, including those related to privacy and online harms.' (p43) However, in fact, the DPSI No 5 Report recognised existing processes and raised caution about the need to 'carefully' consider the 'potential for overlap or conflict with other existing avenues for redress (for example, in relation to online harms, advertising standards, privacy, disinformation and misinformation)'.[48] Neither the DPSI No 5 Report nor Government's statement in response to this report contains sufficient detail, particularly none relevant to online safety, to suggest its relevance to the OSA and therefore to the Statutory Review.

With the DPSI No 5 Report's caution in mind, and as existing work is already underway within industry and government to address these recommendations which have cross-portfolio relevance and impacts, we consider that the issue of dispute resolution processes should be carved out of any recommendations by the Statutory Review.

Regulatory enforcement

With respect to the questions that relate to the enforceability of the OSA framework that are outlined in ***Part 4*** of the Issues Paper, we recognise that effective regulation must be underpinned by a robust enforcement framework. Under the OSA, the eSafety Commissioner has a range of different information gathering and investigatory powers and can pursue a range of different penalties. We consider that the combination of graduated enforcement options, coupled with the transparency mechanisms of the BOSE regime, create a solid foundation for enforcement of the OSA.

[46] See, for example, 'Meta's Quarterly Updates on the Oversight Board', *Transparency Center*, https://transparency.meta.com/en-gb/oversight/meta-quarterly-updates-on-the-oversight-board/

[47] Treasury, *Government response to ACCC Digital Platform Services Inquiry*, 8 December 2023, https://treasury.gov.au/sites/default/files/2023-12/p2023-474029.pdf

[48] ACCC, *Digital platform services inquiry - Interim report No. 5 – Regulatory reform*, September 2022, [4.3.2] https://www.accc.gov.au/system/files/Digital%20platform%20services%20inquiry%20-%20September%202022%20interim%20report.pdf

The eSafety Commissioner's enforcement powers range from formal warnings to infringement notices to civil penalties to injunctions. This broad range of powers enables the Commissioner to adopt a graduated and proportionate approach to enforcement, with more severe penalties being reserved for more severe breaches. In the most severe circumstances, the Commissioner may apply to the Federal Court to obtain an order that a platform cease providing their service in Australia. To date, these powers remain relatively untested, and it is not clear that additional powers are required.

We acknowledge the concerns expressed in the Issues Paper about enforceability of penalties against companies based overseas. However, given the compliance-first, informal approach that has been developed since the OSA's first enactment in 2015, there is no evidence that the fact that a company is based overseas is a significant impediment to compliance or enforcement. For example, Meta has developed compliance measures to address the Phase 1 Codes, and is commencing that process in relation to the Phase 1 Standards, and has responded to multiple rounds of non-periodic reporting notices from the eSafety Commissioner in relation to the BOSE. This compliance-first approach, encouraged by the graduated penalties regime, is also evinced by the relatively small number of formal removal notices that have been issued by the Commissioner, with the Commissioner instead being able to rely on informal cooperation by service providers. To this end, we support eSafety's position that, in many circumstances, informal or less intrusive enforcement action can be far more effective at securing the desired regulatory result.[49]

In addition to the Commissioner's enforcement powers, the Commissioner also has an array of information gathering powers under the OSA. These include the ability to investigate suspected breaches of the OSA, to obtain identity information about a user and to require transparency reporting in relation to the BOSE. While we support the Commissioner's objective of improving the transparency and accountability of industry through the BOSE regime, it is not clear what further information could be obtained through an expansion of such powers.

In particular, under the BOSE transparency powers, the Commissioner is able to seek extensive information and data about a platform's measures, processes and decisions. This makes the Commissioner's current information gathering powers the widest and most robust in the region, and some of the most expansive in the world. In fact, there are very few constraints on the nature of information that may be sought by the Commissioner and very few protections for the confidentiality of such information. This is problematic as the publication of certain information provided in response to a notice could have serious impacts on the safety and integrity of our services. While we welcome eSafety's constructive and cooperative approach to these issues, we consider that the OSA could benefit from formalising a process for the protection of highly

[49] See eSafety Commissioner, *Compliance and Enforcement Policy*, December 2021, p6, https://www.esafety.gov.au/sites/default/files/2022-03/Compliance%20and%20Enforcement%20Policy.pdf?v=1719917281497

sensitive and confidential information. We support accountability through transparency but consider that it should be balanced against the importance of due process.

International approaches to address online harms

While Australia has sought to be a first mover in online safety regulation, given the pace of regulatory reforms in other jurisdictions, we recognise the opportunity presented by the Statutory Review to consider what, if any, elements of international approaches should be considered for introduction locally. The Issues Paper considers measures such as statutory duties, transparency and access to data in ***Questions 21-24*** in Part 5. (note with respect to ***Question 25*** that relates to dispute resolution, we have discussed this above).

It is also important to note that the online safety regulatory regimes in other countries, like the ones in the EU, Ireland and Singapore, are new and still being tested. The implementation of these laws is an iterative process, in which platforms and regulators are all still assessing and learning. As such, it is too soon to know if these laws are working as intended and achieving the regulatory intent. It will be important to conduct targeted regulatory impact assessments of these laws once they are fully implemented to understand their impact.

Statutory Duty of Care

Meta agrees with the underlying principle that service providers should be required to implement measures to make their services safe for users and to minimise harms. As outlined above, we are committed to building our services with safety in mind.

The question is whether a single overarching duty of care or more specific, articulated duties is more appropriate, both to incentivise investments in safety and integrity and to enhance user safety. In our experience of investing significantly in safety and integrity systems, there are significant challenges associated with an overarching, but undefined 'duty of care'. This is because it is uncertain, especially as to where liability might lie, and it creates significant compliance challenges. As an ecosystem, industry and users alike benefit from clarity, certainty and predictability of online safety duties or obligations.

By way of example, the UK's Online Safety Act, has a number of specific and separate duties that are defined in the Act, with definitions in the Act and accompanied codes of practice which further define what actions covered services are required to do to fulfil such duties. This is a model already functionally in place under the current Australian OSA and companion industry codes and standards.

We see two main benefits to maintaining the specific duties approach that maps to the current Australian approach: it enables a proportionate and risk-based regulatory response to be adopted, with stricter compliance requirements being imposed for more serious types of online harm; and it provides greater certainty to users, regulators and industry, with tailored codesgiving service providers clear direction as to what actions are expected of them in relation to each duty, and with a single regulating body who sets the parameters.

Access to data for research

When considering appropriate transparency measures with respect to Australia's online safety framework, we encourage the Statutory Review to consider what work is already being undertaken to provide transparency, data and research by industry. The Issues Paper (and in particular the framing of ***Question 24***) seems to assume that without regulation, there is no transparency and access to data for research. This is not the case.

For many years now, Meta has provided transparency via Meta's Transparency Center.[50] Our Transparency Center provides a one stop-shop that contains details of our policies, enforcement and integrity insights, including in relation to the use of AI to inform ranking of content, our efforts to reduce problematic content and our AI-driven integrity efforts as part of our content governance.[51] We also provide a deeper look at the types of signals and prediction models that we use in our ranking systems to reduce problematic content.[52] And finally, the Transparency Center houses our Community Standards Enforcement Report that provides data on how much harmful content we action, prevalence of harmful content, proactive detection rates as well as appealed and restored content.[53]

In addition, Meta is committed to supporting independent research that will enhance our understanding of the impact our services can have on society. We are also deeply committed to protecting our users' privacy and maintaining a safe and secure community. Meta also recognises the importance of being transparent and sharing meaningful data with researchers - data that is robust, representative and thereby can serve as a basis to contribute to understanding how our services work and their potential impact on society.

Over the years, we have worked to promote research - while preserving privacy - through multiple initiatives. Below are some examples of our work to support independent research:

- A publicly accessible Ad Library and an Ad Library API.[54]

[50] Meta, *Transparency Center*, https://transparency.meta.com/en-gb/

[51] Meta, 'Our approach to ranking explained', *Transparency Center*, June 2023, https://transparency.fb.com/features/explaining-ranking/

[52] Meta, 'Our approach to Facebook Feed ranking', *Transparency Center*, June 2023, https://transparency.fb.com/en-gb/features/ranking-and-content/

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

[54] Meta, Ad Library and Ad Library API, *Transparency Center*, https://transparency.meta.com/en-gb/researchtools/ad-library-tools/

- A Content Library and Content Library API that provide means for researchers to conduct research that is in the public interest, with privacy protective measures as well as measures protecting the security of the Facebook and Instagram services.[55]
- Access to specific datasets for research purposes, including:
  - The Data for Good program to empower partners with data to help make progress on major social issues.[56]
  - Sharing information with independent researchers about our network disruptions relating to Coordinated Inauthentic Behaviour (CIB).[57]
  - URL Shares or Ad Targeting Datasets, the first providing individual-level counts of the number of people who viewed, commented, shared or reacted to an URL, while the second giving access to granular, ad-level targeting information for social issues, electoral and political ads run across Meta's platforms in 120+ countries.[58]

Australian researchers are able to apply for access to Meta's research tools and datasets.[59]

Policy initiatives that seek to provide access to data for research need to consider the complicated tradeoffs between privacy, transparency, proportionality, and other values. These initiatives should be guided by clear principles, and these include:

- **Sharing data for research often involves complicated tradeoffs between privacy and other values; proposals intended to facilitate data sharing should focus on ways to protect privacy while unlocking the benefits of data.** In other words, proposals that aim to facilitate research should do so by helping to address the difficult privacy and liability issues that typically attend this data sharing -- e.g., by incentivising the development and use of privacy-enhancing technologies, developing standardised language for data-sharing agreements, and providing safe harbours for good actors and penalties for bad ones. Data shared with researchers should follow minimum privacy and security standards, including minimisation of data shared for a given purpose, application of privacy-enhancing technologies if possible, purpose limitation on how the data may be used, and deletion upon an agreed time after completion of research and sufficient opportunity for replication and reproduction has elapsed. There should also be protections from using or publishing any personal data, or data that could be used to identify individuals.

[55] Meta, Content Library and Content Library API, *Transparency Center*, https://transparency.meta.com/en-gb/researchtools/meta-content-library/
[56] Meta, *Data For Good*, https://dataforgood.facebook.com/
[57] Meta, 'Meta's Threats Disruptions', *Transparency Center*, https://transparency.meta.com/en-gb/metasecurity/threat-reporting/
[58] Meta, *Facebook Open Research and Transpare*ncy, https://fort.fb.com/researcher-datasets
[59] Meta, Research tools and datasets, https://transparency.meta.com/en-gb/researchtools

- **Policymakers should determine who receives data and under what circumstances; but Platforms should retain the ability to determine how third parties receive the data.** Policymakers should decide who should receive data for research and under what circumstances, but platforms should have the ability to determine the technical mechanism through which data can be shared; “backdoor” access (e.g., through scraping) should not be permitted.

- **Research data shared should be relevant to assessing systemic risks**. Requests for access to data should be specific, that they carefully make clear how the data sought will contribute to an intended research outcome, and explain how that research outcome will be relevant to detecting, identifying and understanding systemic risks on a platform, or relevant mitigation measures. Platforms should not be required to share data that would unduly burden the platform to produce, i.e. that the production of the sought data should not require an unreasonable or disproportionate amount of time or resources in relation to the reasonable needs of the research being performed. Further, disclosure of platforms’ trade secrets or confidential business information should be excluded from data sharing.

- **Platforms that exercise reasonable diligence and care in sharing data with a researcher should be protected from liability if the researcher misuses the data**. Platforms that behave responsibly when selecting research partners, e.g., by performing reasonable diligence and putting in place appropriate privacy protections, shouldn’t face liability for downstream misuse of data.

- **Researchers should be held accountable for misusing data they receive from platforms**. Researchers should share responsibility for protecting people’s data. Where they fail to do so, they and/or their institutions should be held liable.

Other online harm types

As part of its review of the OSA, the Statutory Review is tasked with considering a range of specific harm types such as volumetric (pile-on) attacks, technology-facilitated abuse and technology-facilitated gender-based violence, online abuse of public figures and those requiring an online presence as part of their employment, as well as online hate (see questions in ***Part 3*** and also ***Question 27***).

Whilst we recognise the importance of ensuring that the regulatory framework is fit for purpose to ensure that industry is required and the regulator is empowered to enforce Australian community expectations to combat emerging harms, we encourage the Statutory Review to first consider the work of industry to tackle emerging harms, and also test whether the existing protections under the OSA - such as the notice and takedown scheme, the BOSE and the industry codes and standards - already cover them.

In our experience, emerging harm types such as those outlined in the Terms of Reference can be addressed through the existing framework of the OSA or existing laws.

Emerging harm types (except online hate speech)

Meta has policies and regularly enforces on content that falls within the following categories of harms outlined in the Issues Paper: volumetric (pile-on) attacks, technology-facilitated abuse and technology-facilitated gender-based violence, online abuse of public figures. The enforcement of these policies, in our view, meets BOSE expectation 6: to take reasonable steps to ensure the safe use of our services. We have outlined in more detail in the Appendix below about our efforts against these harm types – such as our work to promote women's safety, protect public figures from abuse and mental health and well-being (including suicide, self-harm and eating disorder content).

To ensure that our policies and responses adapt to emerging harm types, 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 of 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. Members of our Product Policy team run a regular 'Policy Forum' to discuss potential changes to our policies and their enforcement. A variety of internal and external subject matter experts participate in these meetings, and hear input from external groups. In keeping with our commitment to greater transparency, the minutes of these meetings are made publicly available.[60] A change log of changes made to each policy area is available within the Community Standards.[61]

We encourage the Statutory Review to consider any proposals with respect to emerging harm types against the principles of proportionality, flexibility and complementarity. It may be challenging for the OSA to have longevity if there is a high degree of specificity in harm types (given the changing nature of technology and people's use of it). In addition, many other regulatory frameworks already address the subject matter and we should rely on those existing safety regulatory frameworks or existing laws in Australia, for example:

[60] Meta, 'Policy Forum Minutes', https://transparency.meta.com/en-gb/policies/improving/policy-forum-minutes/
[61] See, for example, Meta, 'Facebook Community Standards - Child sexual exploitation, abuse and nudity', *Transparency Center*, https://transparency.meta.com/en-gb/policies/community-standards/child-sexual-exploitation-abuse-nudity/

- Some of the topics considered in the Statutory Review's Issues Paper, e.g. recommender systems, and generative AI, have already been addressed by the BOSE Amendments;
- Others e.g. gender-based violence online are also addressed through existing frameworks covering adult cyber-bullying;
- Online hate is an area that not only overlaps with existing protections against anti-discrimination, but also is an area with clear policies in place for much of industry that is seeking to also address at a systems-level and not just on individual pieces of content; and additionally, additional legislative proposals are being considered by other portfolios[62]; and
- Existing criminal laws have a place with enforcing and deterring against some egregious online conduct, and should be used to address some harms at the perpetrator level, e.g. gender-based violence, online abuse of public figures, and "post and boast" conduct (which is covered under under existing or proposed State laws[63]).

Ensuring consistency in combating online hate speech

Meta shares the Government's intent to ensure that people who use online services are not subjected to hate speech. We have long-standing policies that prohibit hate speech and have steadily increased our investment in proactive detection technology over the years such as that, for example, in the first quarter of 2024, we proactively detected and actioned 94.7 percent of hate speech content on Facebook and 98.2 percent of hate speech content on Instagram, before people reported it.[64] Whilst we have always removed hate speech when becoming aware of it, the increasing use of AI to identify hate speech has meant that we are able to action it more often before people are exposed to it.

Our progress is due in large part to our recent AI advances in a few areas:

- *Lingual understanding:* the ability to build machine learning classifiers that can analyse the same concept in multiple languages - and learning in one language can improve its performance in others. This is particularly useful for languages that are less common on the internet.

[62]See, for example, The Hon Mark Dreyfus KC MP, 'Media Conference – Parliament House', 13 February 2024, https://ministers.ag.gov.au/media-centre/transcripts/media-conference-parliament-house-13-02-2024

[63] See, for example, *Crimes Act 1900* (NSW), s 154K; *Bail Act 2013* (NSW), s 22C; *Criminal Code* (Qld), s 408A.

[64] Meta, *Community Standards Enforcement Report*, Q1, 2024, https://transparency.fb.com/reports/community-standards-enforcement/hate-speech/facebook

- *Whole post understanding or WPIE*[65]*:* the ability to look at a post in its entirety, whether images, video and text, and look for various policy violations simultaneously instead of having to run multiple different classifiers.

We also use artificial intelligence to prioritise content that needs reviewing, after considering several different factors:

- *Virality:* Content that is potentially violating that's being quickly shared will be given greater weight than content that is getting no shares or views.
- *Severity:* Content that's related to real-world harm such as suicide and self-injury or child exploitation will be prioritised over less harmful types of content such as spam.
- *Likelihood of violating:* Content that has signals which indicate that it may be similar to other content that violated our policies will be prioritised over content which does not appear to have violated our policies previously.

Prioritising content in this way, regardless of when it was shared on our services or whether it was reported by a user or detected by our technology, allows us to get to the highest severity content first.

We note that, with the introduction of the recent BOSE Amendments, hate speech is now explicitly within the scope of the BOSE. There is now an expectation that providers will implement processes for detecting and addressing hate speech which breaches a service's terms of use. We consider this to be a balanced and appropriate approach to the regulation of hate speech, which is both hard to define and hard to detect, due to its highly contextual nature. We would caution against the introduction of further reforms to the OSA on this issue, especially in light of the Government's proposal to introduce separate hate speech laws. We encourage the Statutory Review to wait for the outcome of these laws before making any further changes to the OSA in order to avoid unnecessary regulatory complexity and overlap.

[65] 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/

Specific trends and emerging technologies

The Terms of Reference invites consideration as to whether additional arrangements are needed with respect to new and emerging technologies such as generative AI, immersive technologies, recommender systems, and encryption. The Issues Paper specifically invites feedback on the OSA framework in the context of new and emerging technologies (specifically ***Questions 28 and 29***).

When designing regulatory approaches, the challenge for policy makers is deciding between adaptive and flexible frameworks as distinct from those that are prescriptive, and thus reactive rather than forward-looking. We consider that the OSA framework strikes an appropriate balance between both of these approaches. It draws on the benefits of adaptive and flexible systems-based regulation through the BOSE and the industry codes and standards, as well as the benefits of reactive, yet targeted enforcement, through the complaints-handling scheme.

The strength of the OSA as an adaptable regulatory instrument is particularly evident in the way that it regulates the full spectrum of online services in Australia. For example, even though something as new and developing as generative AI is not *explicitly* covered by the OSA, it is still caught by the OSA (e.g. as a “designated internet service”) and regulated through the BOSE and the industry codes and standards. In fact, the breadth of the OSA means that it likely already regulates all of the new and emerging technologies identified by the Issues Paper.

- Encrypted messaging services and virtual reality games are covered by the definition of “relevant electronic services”. These services are regulated by both the BOSE and Phase 1 Standards. Specific provisions around encryption are included in the BOSE.

- Generative AI websites are covered by the definition of “designated internet services”, which are also regulated by the BOSE and Phase 1 Standards. Specific provisions around generative AI are included in both the BOSE and Phase 1 Standards.

- Recommender systems or algorithms used by a social media service will form part of the “social media service” and will be regulated by the BOSE and Phase 1 Codes. Specific provisions around recommender systems are included in the BOSE.

Additionally, the Phase 2 Codes that are now being developed will be required to address the issue of age assurance as one of the key measures designed to prevent children in Australia from accessing or being exposed to class 1C and class 2 material. In line with this, and noting the Statutory Review Issues Paper’s attention to various trends and emerging technologies, we recommend that the Statutory Review focus on the opportunity for Australia to craft a leading global approach in the broader OSA framework with respect to age assurance solutions.

To assist the Statutory Review, we outline below the work that Meta is already undertaking to address policy maker concerns and questions with respect to online safety in the context of these emerging technologies. We welcome the opportunity to bring the insights from our investments and work as part of our compliance with the BOSE, Codes and Standards.

Before turning to our work to address policy maker concerns with respect to technologies such as generative AI, algorithms and recommender systems and encryption, we first detail our recommendations for taking an ecosystem approach to age assurance within the broader OSA framework, while also sharing with the Statutory Review details on our approach and efforts around the other key areas of generative AI, recommender systems and algorithms, as well as end-to-end encryption.

Age assurance solutions

In relation to age assurance measures, the Statutory Review is timely. Australia can lead the way in developing a global approach with respect to age assurance solutions. Debates around ensuring age-appropriate experiences online are occurring globally but the most effective solutions are not yet agreed on.

We have been committed to providing safer, more private, age-appropriate experiences to teens when they use our services, as our submission has laid out. We need privacy-protective solutions that allow the tech industry to provide age-appropriate experiences and parents to be more involved in the apps their teens use. Age verification and parental oversight are a cornerstone of developing youth legislation, but there are no globally accepted industry standards that dictate how technology companies should approach age assurance online. Current laws around the world lack clear guidance on how companies and developers should approach age verification in a compliant, privacy-preserving way. Ecosystem approaches with industry-wide solutions, where all apps are held to the same and consistent standard, would provide the necessary consistency to most effectively protect young people. If the whole industry works together to provide safe, age-appropriate experiences, we can move closer to helping young people avoid flocking to apps or services that are far less safe than those that have invested in age-appropriate protections and experiences.

Understanding a user's real age is key to all these efforts, as it is for all app and service providers seeking to provide age-appropriate experiences. This information allows us to create new safety features for young people, and helps ensure we provide the right experiences to the right age group. As outlined above, in addition to asking for age at sign-up, we do a number of things to assure we have a user's correct age, including the use of AI technology to identify users that may be misrepresenting their age, so that we can take steps to ensure they have age-appropriate experiences, which may include requiring them to verify their age.

Understanding age is an industry-wide challenge because people may misrepresent how old they are, to access apps and services that weren't designed for them. Given this, we have learnt over the years that a multilayered approach is the best path forward for age assurance. Looking across the industry, many apps or services will not have the same resources as currently popular apps or app store operators, and may implement less reliable age assurance measures. If we are to meet the needs of parents and teens, we will then have to aim for solutions that appropriately balance between:

- **Industry wide:** consistency across all the apps and services teens use;
- **Privacy:** minimise the additional data being collected to verify age;
- **Effectiveness:** ensure that methods are sufficiently reliable and hard to circumvent;
- **Fairness:** ensure that methods used to assure age are accessible by diverse global populations.

As has already been proposed by the eSafety Commissioner in the ongoing discussion of the Phase 2 Codes, we advocate for an ecosystem approach to age assurance, with responsibilities and obligations taken by all involved from the operating system / device level down to individual apps and services. We believe that any additional requirements around age assurance should consider the broader ecosystem and require that this also be done at the operating system / device level, where age collection and confirmation data is processed once and thereafter shared across industry throughout the respective app ecosystems.

Parents should have an easy way to review the apps their children use, and verify their children's ages to ensure they have age-appropriate experiences. Rather than having to consider solutions within the hundreds of apps available, and thousands more to be developed in the years to come, we should endeavour to first meet parents and teens where they are at - on their devices. Device and app store solutions simplify parents' oversight and management of their teens' online experiences and provide them with greater comfort knowing that their teens' ages are accurately reflected across the apps they use and services they access.

This simple approach has many benefits; in addition to reducing the onus on parents to find and navigate multiple age verification systems across multiple apps, it also minimises the number of places and times people have to share potentially sensitive data to verify age, allowing us to also prioritise data privacy and security. Today, teens and parents already provide app store operators like Apple and Google with this information when they purchase their devices and set up their accounts, with systems already built by these companies for parental notification, review, and approval into their app stores. We should leverage these systems.

Such measures at the operating system / device level will be complementary to, and should not replace the efforts and responsibilities that services like Meta are already pursuing, yet benefiting new entrants to app ecosystems in ensuring age appropriate experiences for all. On this basis, apps and services can take additional measures to directly verify age when they have signals a user may be misrepresenting their age, and also deploy additional ongoing measures to shore up age assurance.

We recommend that the Statutory Review focus on the opportunity to set a global best practice around age-assurance within the broader OSA framework, together with the expertise of the eSafety Commissioner.

We also provide further detail about our work to provide an age appropriate experience on our services in the Appendix below, including in relation to supervisory tools for parents, how we understand a user's age and our age appropriate controls.

Meta's approach to Generative AI

We also note that the Statutory Review's Issues Paper has also raised generative AI as an emerging technology for consideration within the OSA framework.

There is significant action at an international level to consider the best ways to establish governance frameworks and safety evaluations for AI. This includes the White House Voluntary AI Commitments, the White House *Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*,[66] the Group of Seven (G7) Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, the *Bletchley Declaration* arising out of the UK AI Safety Summit[67], the recent Korean AI Safety Summit, the UN High Level Advisory Board on AI and the UN Global Digital Compact. These complement existing global frameworks, such as the OECD *Principles on Artificial Intelligence* adopted in May 2019 by OECD member countries.[68]

We have participated in many of the international AI governance engagements, specifically, we supported the G7 Hiroshima AI Process,[69] participated in the Bletchley UK AI Summit[70] and

[66] US National Archives Federal Register, 'Safe, Secure, and Trustworthy Development and Use of Artificial Inte*lligence', Executive Order 14110*, 88 FR 75191, 30 October 2023, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[67] UK Government, *'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023'*, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[68] OECD, *'Artificial Intelligence: OECD Principles'*, https://www.oecd.org/digital/artificial-intelligence

[69] https://www.threads.net/@nickclegg/post/CzCimNDPs5_

[70] https://www.threads.net/@nickclegg/post/CzESxRHLHTA

signed the Seoul Frontier AI Safety Commitments[71]. As AI Safety Institutes[72] are established and commence their work in the US, UK and Japan, among others, we support future work plans to undertake effective evaluation of models. We have also publicly committed to Child Safety Generative AI Principles developed by Thorn and All Tech Is Human.[73]

In addition to the considerable work that is already taking place with respect to governance models for GenerativeAI, we would also encourage the Statutory Review to also consider that AI is both a sword and a shield. Meta's Community Standards apply to all content posted on our platforms regardless of how it is created. When it comes to harmful content, the most important thing is that we are able to catch it and take action regardless of whether or not it has been generated using AI. The use of AI in our integrity systems is a big part of what makes it possible for us to catch it.

We have used AI systems to help protect our users for a number of years. For example, we use AI to help us detect and address hate speech and other content that violates our policies. This is a big part of the reason why we have been able to cut the prevalence of hate speech on our platforms within the last few years (for example, from 0.10 to 0.11 percent in Q3 2020, to about 0.02% in Q1 2024.[74] In other words, for every 10,000 content views, we estimate just one or two will contain hate speech.

Meta has been a pioneer in AI development for more than a decade. We know that progress and responsibility can and must go hand in hand. Generative AI tools offer huge opportunities, and we believe that it is both possible and necessary for these technologies to be developed in a transparent and accountable way. That is why we have developed policies and tools to promote the integrity and awareness of the use of AI in content generation.

We have developed specific features for generative AI content, including to incorporate feedback on our handling of manipulated media on Facebook, Instagram and Threads from the Oversight Board and extensive public opinion surveys and consultations with academics, civil society organisations and others as part of our policy review process.[75] For content that does not violate our policies, we believe it is important for people to know when photorealistic content they are seeing has been created using AI. We already label photorealistic images created using Meta AI, and this year have begun labelling a wider range of video, audio and

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

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

[73] See https://www.thorn.org/blog/generative-ai-principles/

[74] Meta, 'Community Standards Enforcement Report - Hate speech', *Transparency Center*, https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook

[75] Meta, 'Our Approach to Labeling AI-Generated Content and Manipulated Media', *Newsroom*, 5 April 2024, https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media

image content when we detect industry standard AI image indicators or when people disclose that they're uploading AI-generated content.[76]

When photorealistic images are created using our Meta AI feature, we take several measures to help people know whether those images are generated by AI, including putting visible watermarks on the images, and both invisible watermarks and metadata embedded within image files. Using invisible watermarking and metadata in this way improves both the robustness of these disclosure mechanisms and helps other platforms identify AI-generated images.

Additionally, we have begun providing a wider range of video, audio and image content with "AI Info" labels when we detect industry standard AI image indicators or when people disclose that they're uploading AI-generated content.[77] If we determine that digitally created or altered image, video or audio content creates a particularly high risk of materially deceiving the public on a matter of importance, we may add a more prominent label, so people have more information and context. Advertisers who run ads related to social issues, elections or politics with Meta also have to disclose if they use a photorealistic image or video, or realistic sounding audio, that has been created or altered digitally, including with AI, in certain cases.[78] AI-generated content is also eligible to be fact-checked by our independent fact-checking partners and we label debunked content so people have accurate information when they encounter similar content across the internet.

We have also supported independent AI ethics research that takes local traditional knowledge and regionally diverse perspectives into account. In 2020, we invested in eight independent research projects in the Asia-Pacific through our Ethics in AI Research Initiative for the Asia Pacific, with award recipients including Monash University and Macquarie University.[79] Continued research and collaboration with experts can assist in supporting technical work that enables AI to be more explainable and predictable.

[76] Meta, 'How Meta Is Preparing for the EU's 2024 Parliament Elections', *Newsroom*, 25 February 2024, https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/; Meta, 'Our Approach to Labeling AI-Generated Content and Manipulated Media', *Newsroom*, 5 April 2024, https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media

[77] Meta, 'How Meta Is Preparing for the EU's 2024 Parliament Elections', *Newsroom*, 25 February 2024, https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections; Meta, 'Our Approach to Labeling AI-Generated Content and Manipulated Media', *Newsroom*, 5 April 2024, https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media

[78] Meta, 'Helping people understand when AI or digital methods are used in political or social issue ads', https://www.facebook.com/government-nonprofits/blog/political-ads-ai-disclosure-policy

[79] Meta, 'Facebook announces award recipients of the ethics in AI research initiative for the Asia-Pacific', *Meta Research Blog*, 18 June 2020, https://research.facebook.com/blog/2020/06/facebook-announces-award-recipients-of-the-ethics-in-ai-research-initiative-for-the-asia-pacific

Recommender systems and algorithms

The Terms of Reference for the Statutory Review invites consideration as to whether additional arrangements are necessary to address potential safety concerns with respect to recommender systems and algorithms. We note that, with the introduction of the BOSE Amendments, recommender systems are now explicitly covered by the BOSE.

To provide the Statutory Review with more context on the role of this technology, and the significant transparency and integrity work that has been undertaken in relation to these, we wanted to share some details about the role of algorithms on Facebook and Instagram and the work to deepen the understanding by policy makers about their functionality, as well as end user controls. The overview of our work on mental health and well-being provided in the Appendix is also helpful to consider in the context of this discussion.

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

We understand there is concern about the role of algorithms and AI in ranking and recommending content. This is why we prioritise providing greater transparency to help users better understand how our ranking algorithms and AI-powered products work and when they are engaging with AI-generated content, as well as provide users with more tools to control what they see in their Feed. Our President of Global Affairs, Nick Clegg, outlined Meta's approach on this in an article published last year.[80] People who use our products should have meaningful transparency and control around how data about them is collected and used, and this should be explained in a way that is understandable. That is why we are:

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

At Meta, we use a range of different algorithms to help us rank content. The ones that people are often most familiar with are those that we use to rank content in their Feeds on Facebook and Instagram. Those algorithms that help with ranking play different roles. Some help us find and remove content from our platform that violates our Community Standards, or filter content that is potentially problematic or sensitive. Others help us understand what content is most

[80] Meta, 'How AI Influences What You See on Facebook and Instagram', *Newsroom*, 29 June 2023, https://about.fb.com/news/2023/06/how-ai-ranks-content-on-facebook-and-instagram/

meaningful to people so we can order it accordingly in their feeds. Below, we have outlined more information about these ranking algorithms as well as some of the algorithms we use to recommend new experiences to people.

It is important to bear in mind that the content people see in their Feeds is not solely due to algorithms: what people see is heavily influenced by their own choices and actions. Content ranking is a dynamic partnership between people and algorithms.Even though the people that use our services play a significant role in the ranking process, we recognise that they are only going to feel comfortable with these algorithmic systems if they have more visibility into how they work and then have the ability to exercise more informed control over them. That is why we have been releasing products, tools and greater transparency about the way algorithms work on our services. Our Content Distribution Guidelines[81] and Recommendation Guidelines,[82] explained in more detail below, both set a higher benchmark than our Community Standards; they apply to content that would not otherwise violate our rules on Facebook and Instagram.

Role of algorithms

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

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

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

Historically, these feeds showed content in chronological order, but as more people started using our services and more content was shared, it was impossible for people to see all of the content that was shared, much less the content that they cared about. Instagram, for example, launched in 2010 with a chronological feed but by 2016, people were missing 70 per cent of all their posts in Feed, including almost half of posts from their close connections. So we developed and introduced a Feed that ranked posts based on what people cared about most.[83] Similarly, on Facebook, the goal of Feed is to arrange the posts from friends, Groups and Pages people follow

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

[82] Facebook, ‘What are recommendations on Facebook?’, *Help Centre*, https://www.facebook.com/help/1257205004624246; Instagram, ‘What are recommendations on Instagram?’, *Help Centre*, https://help.instagram.com/313829416281232

[83] Meta, ‘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

to show what matters most at the top of their feed. Our ranking algorithms use thousands of signals to rank posts for each person's Feed with this goal in mind.[84] As a result, each person's Feed is highly personalised and specific to them. Our ranking system personalises the content for over a billion people and aims to show each of them content we hope is most valuable and meaningful, every time they come to Facebook or Instagram.

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

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

One way we measure whether something creates long-term value for a person is to ask them. For example, we survey people[85] 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.[86] Then we can take each prediction into account for a person based on what people tell us (via surveys) is worth their time.

While a post's engagement — or whether people like it, comment on it, or share it — can be a helpful indicator that it is interesting to people, this survey-driven approach, which largely occurs outside the immediate reaction to a post, gives a more complete picture of the types of posts people find most valuable, and what kind of content detracts from their Feed experience. We are continuously working on building out these surveys by asking new questions about the content people find valuable, and we have made it much easier for people to tell us what content they do not enjoy seeing in their Feed.[87]

[84] Meta, 'How does News Feed predict what you want to see?', *Newsroom,* 26 January 2021, https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see
[85] Meta, 'Using surveys to make News Feed more personal', *Newsroom,* 16 May 2019, https://about.fb.com/news/2019/05/more-personalized-experiences
[86] Meta, 'How users help shape Facebook', *Newsroom,* 13 July 2018, https://about.fb.com/news/2018/07/how-users-help-shape-facebook/
[87] Meta, 'Incorporating more feedback into News Feed ranking', *Newsroom,* 22 April 2021, https://about.fb.com/news/2021/04/incorporating-more-feedback-into-news-feed-ranking/

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

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

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

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

Transparency of content distribution policies & tools

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

- *Why Am I Seeing this post?* - helps users to better understand and more easily control what they see from friends, Pages and Groups in their Feed. Users are able to tap on posts and ads in Feed, get context on why they are appearing (such as how their past interactions impact the ranking of posts in their Feed), and take action to further personalise what they see.[88] This includes the ability to customise their Feed, such as switching between an algorithmically-ranked Feed and a feed sorted chronologically with the newest posts first,[89] as well as indicating if they are interested or not interested in the post to inform future content recommendations.

- *Why Am I seeing this Ad?* - provides users with context on their ads, to help them understand how factors like basic demographic details, interests and website visits contribute to the ads in their Feed. We are continually improving our transparency offerings to reflect feedback we receive. In 2023, we updated this tool to provide users with clear information about the machine learning models that help determine the ads they see on Facebook and Instagram Feed.[90]

- *Ad Preferences* - allows users to adjust the ads they see while on Facebook and gives them the ability to update their ad settings to control information we can use to show their ads.[91]

- *Control what you see on Facebook and Instagram* - helps users to learn more about and control what kind of posts they may see on Facebook and Instagram, including who they see posts from.[92]

- *Content recommendation controls* - our content recommendation controls - known as “Sensitive Content Control” on Instagram and “Reduce” on Facebook – allows people to filter more potentially sensitive content or accounts from places like Search and Explore.[93]

As well as providing transparency at the user level, we recognise that there continue to be discussions about the best ways to provide model and systems documentation that enables meaningful transparency around how these systems are trained and operate. Our transparency initiatives at system level include the release of more than 22 AI System Cards that explain how the AI systems in our products work.[94] They give information, for example, about how our AI

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

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

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

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

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

[93] Meta, ‘Introducing Sensitive Content Control’, *Newsroom*, 20 July 2021, https://about.fb.com/news/2021/07/introducing-sensitive-content-control; Facebook, ‘Manage how content ranks in your Feed using Reduce’, *Help Center*, https://www.facebook.com/help/543114717778091

[94] Meta Resources, System Cards, https://ai.meta.com/tools/system-cards/

systems rank content, some of the predictions each system makes to determine what content might be most relevant to users, as well as the controls users can use to help customise their experience.

This is complemented by Meta's Transparency Center[95] and Privacy Center[96]. Our Transparency Center provides a one stop-shop that contains details of our policies, enforcement and integrity insights, including in relation to the use of AI to inform ranking of content, our efforts to reduce problematic content and our AI-driven integrity efforts as part of our content governance.[97] The Privacy Center also includes guides about Generative AI and a Teen Generative AI Guide.[98] We also provide a deeper look at the types of signals and prediction models that we use in our ranking systems to reduce problematic content.[99] And finally, the Transparency Center houses our Community Standards Enforcement Report that provides data on how much harmful content we action, prevalence of harmful content, proactive detection rates as well as appealed and restored content.[100]

Additionally, our Privacy Centre informs people of how we build privacy into our products, including how we use information for generative AI models and features,[101] and how users can manage and control their privacy on Facebook, Instagram, Messenger and other Meta products. This includes instructions to change or delete their information from chats with AIs from Meta,[102] and Meta support and resources for teens relating to generative AI.[103]

Providing guidelines for ranking

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

As mentioned above, we have published Facebook's Content Distribution Guidelines to share more detail on the types of content that we demote in Feed[104], and likewise for Instagram Feed and Stories. While the Community Standards make it clear what content is removed from our

[95] Meta, *Transparency Center*, https://transparency.meta.com/en-gb/
[96] Meta, *Privacy Center*, https://www.facebook.com/privacy/center
[97] Meta, 'Our approach to ranking explained', *Transparency Center*, June 2023, https://transparency.fb.com/features/explaining-ranking/
[98] See information about GenAI: https://www.facebook.com/privacy/genai; a Generative AI Guide: https://www.facebook.com/privacy/guide/generative-ai/ and the Teen Generative AI Guide: https://www.facebook.com/privacy/dialog/an-introduction-to-generative-ai-teens
[99] Meta, 'Our approach to Facebook Feed ranking', *Transparency Center*, June 2023, https://transparency.fb.com/en-gb/features/ranking-and-content/
[100] Meta, Community Standards Enforcement Report, *Transparency Center*, https://transparency.fb.com/data/community-standards-enforcement/
[101] Meta, 'How Meta uses information for generative AI models and features', *Privacy Center*, https://www.facebook.com/privacy/genai
[102] Meta, 'Generative AI at Meta', *Privacy Center*, https://www.facebook.com/privacy/guide/generative-ai/
[103] Meta, 'Access support and resources for teens', *Privacy Center*, https://www.facebook.com/privacy/guide/teens/
[104] Meta, 'Types of content we demote', *Transparency Center,* 20 December 2021, https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/

services because we do not allow it, the Content Distribution Guidelines make it clear what content receives reduced distribution because it is problematic or low quality. Many of these guidelines have been shared in various announcements, but in efforts to make them more accessible, we have brought them together in one easy-to-navigate space in our Transparency Center and Help Center.

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

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

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

Since 2021, we have also published a quarterly Widely Viewed Content Report (WVCR), which aims to provide more transparency and context about what people are seeing on Facebook by sharing the most-viewed domains, links, Pages and posts for a given quarter in Feed in the United States.[105] The WVCR provides additional insights into the different content types that appear in News Feed to help people better understand our distribution systems and how that influences the content people see on our platform. We plan to expand the scope of this report to other countries in future iterations. It will continue to appear in conjunction with our quarterly Community Standards Enforcement Report.

We continually evaluate the effectiveness of Feed ranking signals. We are also making an effort to provide people with more detail about our ranking processes in general. For example, in 2021, the CEO of Instagram published a blog post detailing the ranking process on Instagram from start to finish, which we updated last year.[106]

[105] Meta, 'Widely Viewed Content Report: What People See on Facebook', *Transparency Center*, https://transparency.meta.com/en-gb/data/widely-viewed-content-report/

[106] Meta, 'Shedding more light on how Instagram works', *Instagram Blog*, 8 June 2021, https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works; A Mosseri, 'Instagram Ranking Explained', *Instagram Blog*, 31 May 2023, https://about.instagram.com/blog/announcements/instagram-ranking-explained

Providing guidelines for recommendations

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

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

To help people better understand our approach to recommendations, in August 2020, we published a set of Recommendation Guidelines, which outline the types of content that may not be eligible for recommendations.[107] In developing these guidelines, we consulted 50 leading experts specialising in recommendation systems, expression, safety and digital rights. Recommendation Guidelines are available for both Facebook[108] and Instagram.[109]

End-to end encryption

The Terms of Reference also invite consideration by the Statutory Review as to whether additional arrangements are necessary with respect to encrypted services. We note that the Phase 1 Standards have addressed this and Meta and WhatsApp have both responded to two rounds of reporting notices in relation to the BOSE, which required them to outline the integrity measures specific to their encrypted messaging products].

Consequently, we suggest that the appropriate regulatory guardrails have already been established in Australia with respect to safety mitigations in relation to encrypted services. It is not a matter for appropriate benchmarking across industry to identify best practice.

To assist the Statutory Review with the task before it, we wanted to share more details about Meta's approach to online safety within the context of encryption. Specifically, we apply end-to-end encryption to WhatsApp messages and calls and last year announced the launch of default end-to-end encryption for personal messages and calls on Messenger and Facebook, as well as a suite of new features to let users further control their messaging experience.[110]

[107] Meta, 'Recommendation guidelines', *Newsroom,* 31 August 2020, https://about.fb.com/news/2020/08/recommendation-guidelines/

[108] Facebook, 'What are recommendations on Facebook?', *Help Centre,* https://www.facebook.com/help/1257205004624246

[109] Instagram, 'What are recommendations on Instagram?', *Help Centre,* https://help.instagram.com/313829416281232

[110] Meta, 'Launching Default End-to-End Encryption on Messenger', *Newsroom*, 6 December 2023, https://about.fb.com/news/2023/12/default-end-to-end-encryption-on-messenger/

The extra layer of security provided by end-to-end encryption means that the content of users' messages and calls with friends and family are protected from the moment they leave their device to the moment they reach the receiver's device. This ensures only the user and the person they are communicating with can read or listen to messages, photos, videos, voice messages, and documents sent. No one else can read or listen to that content, not even Meta.

For WhatsApp, we provide a range of features to empower users to keep themselves safe. That includes:

- **Unknown senders:** the first option users are given when someone who is not a contact messages them is whether they would like to block or report them.
- **Block and report:** we advise users to block and report suspicious messages, turn on two step verification for extra security and never click on links or share personal details with someone they do not know. When users choose to report a message, group or other user, that content is reported to WhatsApp for review. Reporting the content means it can be seen by our trust and safety team, who can then pass it onto law enforcement if it is illegal.
- **Privacy settings:** users can adjust their privacy settings to control who sees their information, including their "last seen" and "online", profile photo, about, or status to determine who can see their profile photo, about, or status and who can add them in groups.
- **Mute:** we give users options to mute notifications and archive chats to avoid unwanted interactions. Users can also silence calls from unknown callers.

We encourage users to think carefully before sharing something with their WhatsApp contacts. When a chat, photo, video, file or voice message is shared with someone else on WhatsApp, they will have a copy of these messages and can forward or share with others if they choose to.

As part of our roll out of end-to-end encryption on Messenger, we introduced new privacy, safety and control features, which supplemented existing safety features like report, block and message requests.[111] This includes delivery controls that let people choose who can message them and 'app lock', which uses a device's privacy settings like fingerprint or face authentication to unlock the Messenger app.

We work closely with outside experts, academics, advocates and governments to identify risks and build mitigations to ensure that privacy and safety go hand-in-hand.

[111] Meta, 'Launching Default End-to-End Encryption on Messenger', *Newsroom*, 6 December 2023, https://about.fb.com/news/2023/12/default-end-to-end-encryption-on-messenger/; Meta, 'Messenger Introduces App Lock and New Privacy Settings', *Newsroom*, 22 July 2020, https://about.fb.com/news/2020/07/messenger-app-lock-and-privacy-settings/

Maintaining the integrity of encrypted services

We recognise that there is general agreement across industry, civil society and within Government about the value of encryption to promote privacy, safety, and security. While there are concerns that have been raised about the ability to promote safety and combat harmful content on encrypted services, for Meta, the values of safety, privacy, and security are mutually reinforcing. An independent Human Rights Impact Assessment of Meta's expansion of end-to-end encryption - conducted by NGO Business for Social Responsibility in line with UN Guiding Principles on Business and Human Rights - found, among other areas, that encryption increased the realisation of privacy, freedom of expression, protection against cybercrime threats, physical safety, freedom of belief and religious practices and freedom from state-sponsored surveillance and espionage.[112]

In line with these findings, we continue to invest in behavioural analysis and metadata as effective harm prevention rather than undermine encryption, which is what we have done and are continuing to progress. This is consistent with the eSafety Commissioner's position that services 'should invest in innovative approaches to prevent the dissemination of illegal and harmful content and to detect illegal and harmful content – without compromising privacy and security – to ensure user safety on end-to-end services.[113]

[112] Business for Social Responsibility, *Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption*, 2022, https://www.bsr.org/reports/bsr-meta-human-rights-impact-assessment-e2ee-report.pdf

[113] eSafety Commissioner, *'Updated Position Statement - End-to-end encryption'*, October 2023, https://www.esafety.gov.au/sites/default/files/2023-10/End-to-end-encryption-position-statement-oct2023.pdf?v=1720061393590; eSafety Commissioner, 'Statement on end-to-end encryption and draft industry standards', *Media Releases*, 19 December 2023, https://www.esafety.gov.au/newsroom/media-releases/statement-on-end-to-end-encryption-and-draft-industry-standards

Appendix

Supporting young people and parents

Creating an experience on Facebook and Instagram that is safe and private for young people, but also valuable and relevant, comes with competing challenges. In order to make sure we are striking the right balance, we engage closely with experts in this space - and with young people themselves. We have also engaged with parent groups to better understand the resources they need[114].

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

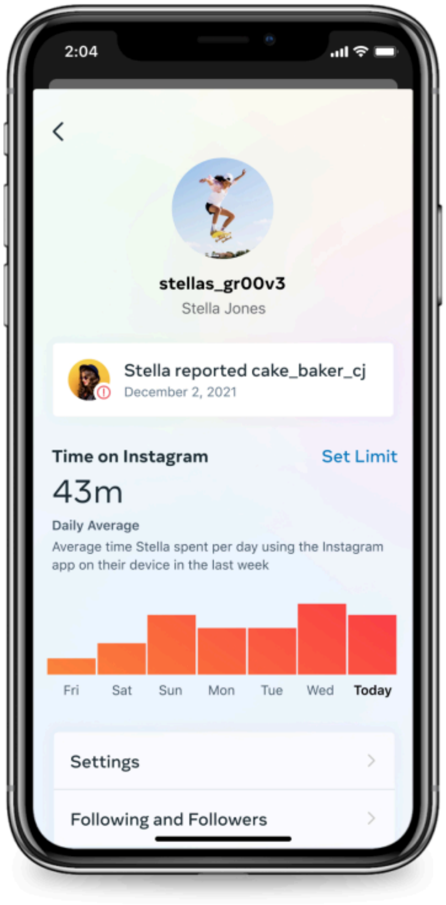
In addition to the responsibility of industry to invest in safety, parents and carers play a vital role in ensuring the safety of young people online. We want to provide tools and resources for parents and guardians so they can guide and support their teens.

Since December 2021, we have launched new tools on Instagram and Facebook that provide greater controls for parents and guardians. Parents and guardians are able to view how much time their teens spend on our platforms and set time limits, shown in Figure 2 below.[115] Teens are also able to notify their parents if they report someone, giving their parents the opportunity to talk about it with them. Parents can also approve or deny changes their teens make to their settings, for example to their account privacy settings.

Figure 2: Parent and guardian controls over 'time spent' and reporting

[114] Meta, 'Introducing Family Centre and Parental Supervision Tools on Instagram and in VR', *Newsroom*, 16 March 2022, https://about.fb.com/news/2022/03/parental-supervision-tools-instagram-vr

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



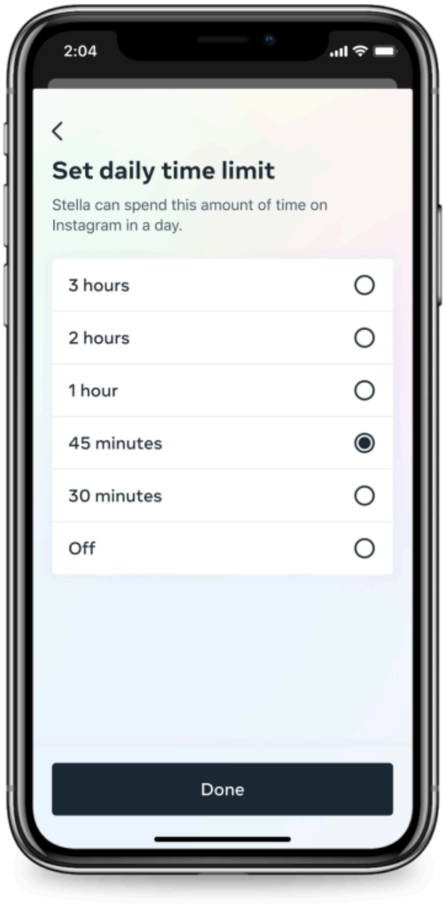
Stella reported cake_baker_cj
December 2, 2021

Time on Instagram

43m

Daily Average

Average time Stella spent per day using the Instagram app on their device in the last week



Set daily time limit

Stella can spend this amount of time on Instagram in a day.

- 3 hours ○
- 2 hours ○
- 1 hour ○
- 45 minutes ●
- 30 minutes ○
- Off ○

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. Examples of these are:

- **Family Centre.** We have developed resources, accessible from within the apps' supervisory experiences, that include product tutorials and tips from experts, to help parents and guardians discuss social media use with their teens.[116]

- **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.[117]

[116] Meta, 'Supporting safer and more positive experiences for your family, *Family Center*, https://familycenter.meta.com/au
[117] Meta, 'Parents', *Safety Center*, https://www.facebook.com/safety/parents

- **Parents' Guide to Instagram.** In Australia, we worked with ReachOut to develop a Parents' Guide to Instagram to support parents in better understanding Instagram's safety tools. The Guide contains tips for parents on using Instagram's safety features and on how to have effective conversations with their teens about social media. The Parents' Guide can be downloaded for free on ReachOut's website and we supported ReachOut to publish the Guide and promote it on their social platforms.[118] The Guide was first released in September 2019 and updated in June 2021.[119] We are working with ReachOut to provide a new series of resources for parents in 2024.

Ensuring age-appropriate experiences online

As per our terms, we require people to be at least 13 years old to sign up for Facebook or Instagram. Our approach to understanding a user's age aims to strike a balance between protecting people's privacy, wellbeing, and freedom of expression.

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

Understanding a user's age

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

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

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. it does not assume that someone is old enough to use our service), and we restrict people who repeatedly try to enter different birthdays into the age screen.

But we also recognise that some people may misrepresent their age online. For that reason, we have been investing in AI tools to help us understand someone's real age. Our technology allows us to estimate people's ages, like if someone is below or above 18, using signals. Technology like this is new, evolving and as many experts in this space have noted, it is not perfect. It also may

[118] ReachOut, *A parents guide to Instagram*, https://parents.au.reachout.com/-/media/parents/files/pdfs/parents_guide_to_instagram_austrlian_edition2021_reachout.pdf

[119] Meta Policy AU, 'A Parent's Guide to Instagram', Meta Australia Policy Blog, *Medium*, 22 June 2021 (updated 27 January 2023), https://medium.com/meta-australia-policy-blog/a-parents-guide-to-instagram-in-partnership-with-reach-out-30a865e28fcb

not always be the most appropriate measure for all use cases. Inaccurate AI predictions could undermine people's ability to use services, for example, by incorrectly blocking them from an app or feature based on false information. Where we do feel we need more information, we have developed a menu of options for someone to prove their age on Instagram and Facebook, which we explain in more detail below.

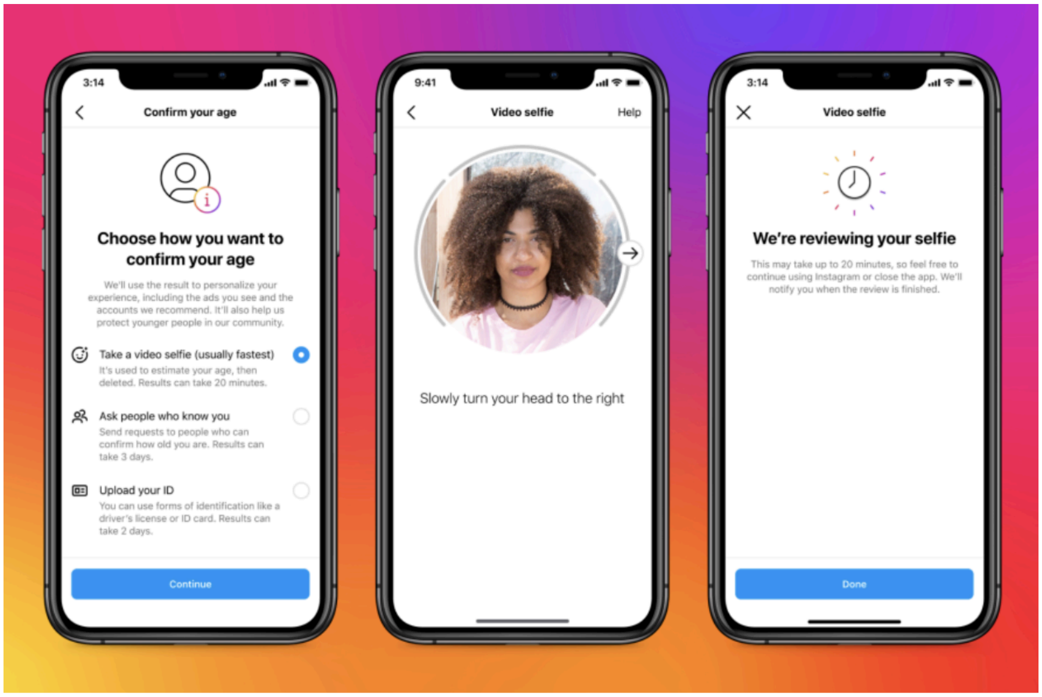
Our team of over 15,000 content reviewers is trained to flag reported accounts that appear to be used by people who are underage. If these people are unable to prove they meet our minimum age requirements, we delete their accounts.

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

We are currently testing two additional options for age verification on Facebook and Instagram,[120] which we have launched internationally, including in Australia: (A) Submission of ID or (B) Face-based-age-prediction, offered through Yoti,[121] a third party vendor based out of the UK which provides age estimation services. This is a new option where users upload a video selfie of themselves to verify their age. Face-based age prediction refers to computer-vision systems, which predicts a user's age based on an image of their face. We partner with Yoti because of their industry leading accuracy metrics, their work to minimise bias across skin tones and gender, and their strong privacy guardrails.

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

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



Choose how you want to confirm your age

We'll use the result to personalize your experience, including the ads you see and the accounts we recommend. It'll also help us protect younger people in our community.

- Take a video selfie (usually fastest) ● It's used to estimate your age, then deleted. Results can take 20 minutes.
- Ask people who know you ○ Send requests to people who can confirm how old you are. Results can take 3 days.
- Upload your ID ○ You can use forms of identification like a driver's license or ID card. Results can take 2 days.

Slowly turn your head to the right

We're reviewing your selfie

This may take up to 20 minutes, so feel free to continue using Instagram or close the app. We'll notify you when the review is finished.

We are also in discussions with the wider technology industry on how best to share information in privacy-preserving ways that helps apps establish whether people are over a specific age.[122] Globally, we believe requiring app stores to get parents' approval whenever their teens under 16 download apps help us place teens in age-appropriate experiences. By verifying a teen's age in the app store, individual apps would not be required to collect potentially sensitive identifying information. Apps would only need age confirmation from the app store to ensure teens are placed in the right experiences for their age group.

Age-appropriate controls

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. We do not allow people who do not mutually follow teens to tag or mention them, or include their content in Reels Remixes. For young people who already have a public

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

account on Instagram, we show them notifications highlighting the benefits of a private account and explaining how to change their privacy settings.[123]

- **Default account limitations.** We also place a range of default limits on a teen's accounts. For example, on Facebook, location sharing is off for minors by default, and we protect certain information, such as minors' contact info, school and birthday, from appearing in search to a public audience.[124] On Instagram and Messenger, adults are unable to start a conversation with a teen who is not connected with them. Over and above this, we prevent teens from messaging potentially suspicious adults (e.g. who may have been blocked or reported by other teens) they are not connected with, and also prevent teens from receiving messages from all accounts they are not connected with by default.[125]

- **Limiting advertisers' ability to reach young people.** We only allow advertisers to target ads to people under 18 based on their age and location. This means that previously available targeting options, like those based on interests or on their activity on other apps and websites, are not available to advertisers. This is in addition to age-gating controls made available for those advertisers who publish age-sensitive ads or content (such as related to gambling).[126]

- **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; and
  - 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.

[123] Instagram, 'Giving Young People a Safer, More Private Experience', *Instagram blog*, 27 July 2021, https://about.instagram.com/blog/announcements/giving-young-people-a-safer-more-private-experience; Instagram, 'About Instagram teen privacy and safety settings', *Help Center*, https://help.instagram.com/3237561506542117/?helpref=popular_articles

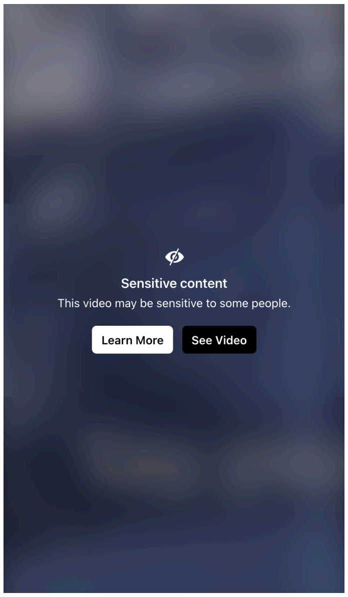
[124] Meta, 'Teen Privacy and Safety Settings', *Meta Policies*, https://www.meta.com/help/policies/safety/Meta-Teen-Privacy-Safety-Settings

[125] Meta, 'Teen Privacy and Safety Settings', *Meta Policies*, https://www.meta.com/help/policies/safety/Meta-Teen-Privacy-Safety-Settings; Instagram, 'About Instagram teen privacy and safety settings', *Help Center*, https://help.instagram.com/3237561506542117

[126] Meta, 'Continuing to Create Age-Appropriate Ad Experiences for Teens', *Newsroom*, 10 January 2023, https://about.fb.com/news/2023/01/age-appropriate-ads-for-teens

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 3 below. The content will not appear, nor allows the option to view, for users who are under the age of 18.

Figure 3: Example of a piece of content that is “marked as sensitive” on Facebook



Sensitive content

This video may be sensitive to some people.

- **Safety Notices in messaging.** In addition to the restrictions explained above, we also send safety notices to users in Messenger and Instagram, if we believe an adult could be pursuing a potentially inappropriate private interaction with a teen.[127]

- **Making it more difficult for adults to find and follow teens.** We have developed new technology that allows us to find accounts that have shown potentially suspicious behaviour and stop those accounts from interacting with young people’s accounts. By “potentially suspicious behaviour”, we mean accounts belonging to adults that may have recently been blocked or reported by a young person, for example.[128]

  Using this technology, we do not show young people’s accounts to these adults who exhibit “potentially suspicious behaviour”. If they find young people’s accounts by

[127] Meta, ‘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; Instagram, ‘About Instagram teen privacy and safety settings’, *Help Center*, https://help.instagram.com/3237561506542117/?helpref=popular_articles

[128] Meta, ‘Giving Young People a Safer, More Private Experience on Instagram’, *Newsroom*, 27 July 2021, https://about.fb.com/news/2021/07/instagram-safe-and-private-for-young-people

searching for their usernames, they will not be able to follow them. They also will not be able to see comments from young people on other people's posts, nor will they be able to leave comments on young people's posts. The reverse also holds - that is, teen accounts' ability to interact with such accounts are similarly restricted. We will continue to look for additional places where we can apply this technology.[129]

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 have 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 and contextual 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 is 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 first quarter of 2024:[130]

- We actioned 7.9 million pieces of content on Facebook for violating our policies on bullying and harassment, and of that, 85.6 percent of bullying and harassment content was removed proactively via AI. This is an increase from 54.1 percent in the first quarter of 2021.

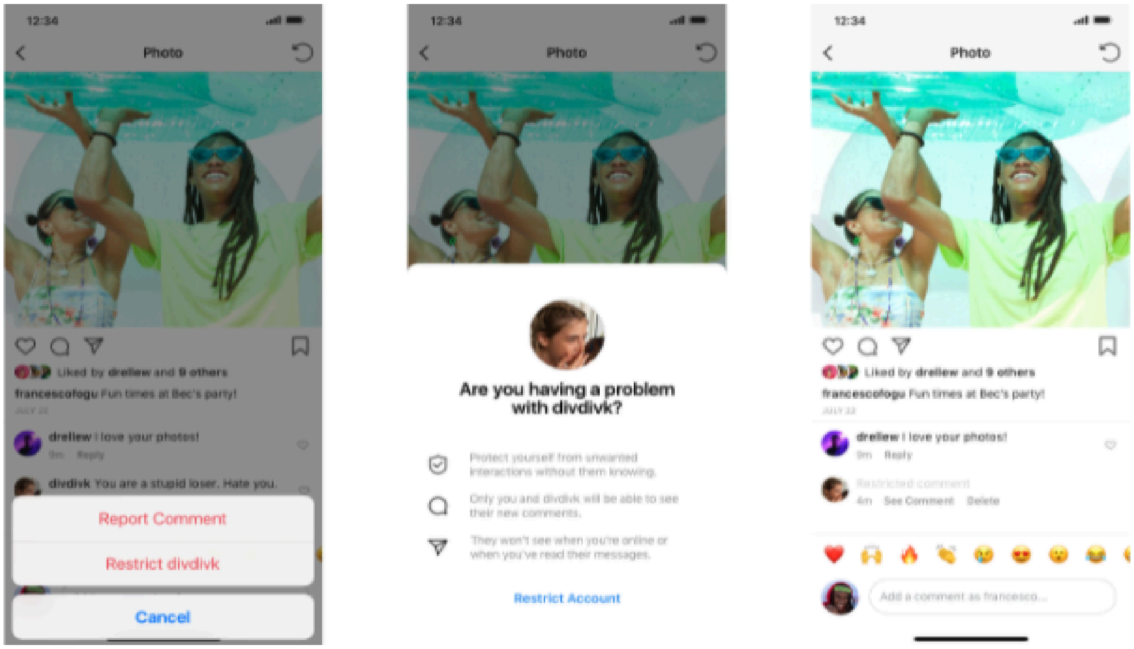
- We actioned 10.3 million pieces of bullying and harassment content on Instagram, and of that, 96.1 percent of it was removed proactively. This is an increase from 78.6 percent in the first quarter of 2021.

[129] Meta, 'Giving Young People a Safer, More Private Experience on Instagram', *Newsroom*, 27 July 2021, https://about.fb.com/news/2021/07/instagram-safe-and-private-for-young-people
[130] Meta, 'Community Standards Enforcement Report Q1 2024 - Bullying and harassment', *Transparency Center,* https://transparency.fb.com/data/community-standards-enforcement/bullying-and-harassment/facebook

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 have invested in a range of other industry-leading tools including:

- **Restrict tool.** We have created a Restrict tool in Instagram[131], shown in Figure 4 below, where comments on a user's posts from a person they have restricted will only be visible to that person. Direct messages will automatically move to a separate Message Requests folder, and the user will not receive notifications from a restricted account. Users can still view the messages but the restricted account will not be able to see when they have read their direct messages or when you are active on Instagram. This feature was developed in direct response to feedback from teens who told us that blocking can be too severe and they wanted a way to protect themselves, but still be able to keep an eye on a bully's activity.

Figure 4: Instagram 'Restrict' tool



12:34 Photo

Liked by drellew and 9 others
francescofogu Fun times at Bec's party!
JULY 22
drellew I love your photos!
9m Reply
divdivk You are a stupid loser. Hate you.

Are you having a problem with divdivk?

Protect yourself from unwanted interactions without them knowing.

Only you and divdivk will be able to see their new comments.

They won't see when you're online or when you've read their messages.

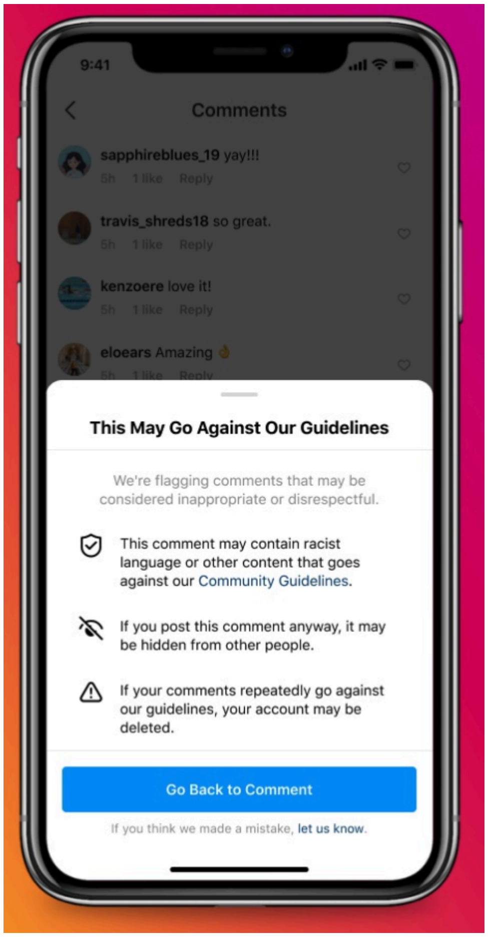
drellew I love your photos!
9m Reply
Restricted comment
4m See Comment Delete
Add a comment as francesco...

- **Bullying and harassment warning**. We send warnings on both Facebook and Instagram to educate and discourage people from posting or commenting in ways that could be bullying and harassment, shown in Figure 5 below. We have found that after viewing these warnings on Instagram, about 50 percent of the time the comment was edited or

[131] 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

deleted by the user.[132]

Figure 5: Warnings to discourage bullying or harassment



sapphireblues_19 yay!!!

travis_shreds18 so great.

kenzoere love it!

eloears Amazing 👌

This May Go Against Our Guidelines

We're flagging comments that may be considered inappropriate or disrespectful.

This comment may contain racist language or other content that goes against our Community Guidelines.

If you post this comment anyway, it may be hidden from other people.

If your comments repeatedly go against our guidelines, your account may be deleted.

If you think we made a mistake, let us know.

- **Limits.** We enable users on Instagram to limit comments, messages, tags and other interactions to existing followers, and turn them off for accounts that do not follow them. When we detect someone might be experiencing a wave of bullying, we show them a notification of this tool asking if they would like to temporarily limit interactions. We recently expanded this in May 2024 to also give people the option of limiting interactions to those on their 'close friends' list only.[133]

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,

[132] Meta, 'Our approach to addressing bullying and harassment', *Newsroom,* 9 November 2021, https://about.fb.com/news/2021/11/how-meta-addresses-bullying-harassment

[133] 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

women can experience a disproportionately higher level of harassment and abuse in the digital world. We care deeply about the issue of violence against women and know that societal issues like misogyny and abuse that occur offline, can also take place online. 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 that their 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 by tailoring 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 women's safety roundtables across the world, including Australia. These roundtables inform our ongoing policy development, tools and programs such as the pilot to address the non-consensual sharing of intimate images (**NCII**) outlined below.[134]

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,[135] 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 focuses on AI-facilitated abuse, deepfakes, gendered violence and image-based sexual abuse. As mentioned above, in Australia, we have also worked closely with WESNET,[136] the Australia national peak body for specialist women's domestic and family violence services, since 2014.

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 of some forms of online harassment and abuse that can occur, especially for women. In July 2019, for example, we expanded our bullying

[134] Meta, 'Making Facebook a safer, more welcoming place for women', *Newsroom,* 29 October 2019, https://about.fb.com/news/2019/10/inside-feed-womens-safety
[135] Meta, 'Partnering with experts to promote women's safety', *Newsroom*, 30 June 2021, https://about.fb.com/news/2021/06/partnering-with-experts-to-promote-womens-safety
[136] WESNET, https://wesnet.org.au

and harassment policy to enforce more strictly on cursing that uses female-gendered terms. Under our hate speech policy, we prohibit direct attacks including violent or dehumanising speech and harmful stereotypes on the basis of protected characteristics including sex, gender identity and sexual orientation.[137] Additionally, under our dangerous organisations and individuals policy, we remove glorification or support of designated individuals or organisations; however, we allow users to share content that mentions them for awareness raising or condemnation.

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.[138] 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.

As mentioned above, through our Recommendation Guidelines,[139] we work to avoid making recommendations that could be low-quality, objectionable, or particularly sensitive, as well as avoiding making recommendations that may be inappropriate for younger viewers. Our Recommendation Guidelines are designed to maintain a higher standard than our Community Standards, because recommended content and connections are from accounts or entities that users have not chosen to follow. Therefore, not all content allowed on our services will be eligible for recommendation. We do not recommend content that may depict violence, or content that may be sexually explicit or suggestive. We generally do not recommend accounts that have recently violated our policies, or are associated with offline movements of organisations that are tied to violence.

Tools

Tools such as blocking, unfollow, reporting, show more/less 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 is already feeling anxious or threatened may not want to trigger a harasser for fear of retribution. Sometimes, the behaviour is not 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

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

[138] Meta, 'Advancing our policies on online bullying and harassment', *Newsroom,* 13 October 2021, https://about.fb.com/news/2021/10/advancing-online-bullying-harassment-policies

[139] Facebook, 'What are recommendations on Facebook?', *Help Centre*, https://www.facebook.com/help/1257205004624246; Instagram, 'What are recommendations on Instagram?', *Help Centre*, https://help.instagram.com/313829416281232

community as friends. That is 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 NCII. It has long been our policy on Facebook and Instagram to remove NCII, and in 2018 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.[140]

Following the success of this pilot, in 2021 we launched the expansion of the program globally, known as StopNCII.org.[141] StopNCII.org operates in partnership with more than 50 non-governmental organisations around the world, including the Office of the eSafety Commissioner.[142]

This is the first global initiative of its kind to safely and securely help people who are concerned that their intimate images (photos or videos of a person which feature nudity or are sexual in nature) may be shared without their consent.[143]

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 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 have 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.

[140] Meta, 'NCII Pilot history', *Safety Center,* https://about.meta.com/actions/safety/topics/bullying-harassment/ncii/pilot

[141] Meta, 'Strengthening Our Efforts Against the Spread of Non-Consensual Intimate Images', *Newsroom,* 2 December 2021, https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images

[142] StopNCII.org - Stop Non-Consensual Intimate Image Abuse, https://stopncii.org

[143] Meta, 'Strengthening our efforts against the spread of non-consensual intimate images', *Newsroom,* 2 December 2021, https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-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 StopNCII.org,[144] which, as detailed above, 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.[145]
- A dedicated safety page for women on our Safety Centre Hub.[146]

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, we provide protections for public figures in recognition of 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, for example updating our policies to increase enforcement against harmful content for public figures.[147] 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.

First, as noted, we have 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 have increased protections for involuntary public figures, such as human rights defenders and journalists. For example, content

[144] Meta, 'Strengthening our efforts against the spread of non-consensual intimate images', *Newsroom,* 2 December 2021, https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images
[145] Meta, *Stop Sextortion Hub*, https://about.meta.com/actions/safety/topics/bullying-harassment/stop-sextortion
[146] Meta, 'Women's Safety', *Safety Center,* https://www.facebook.com/safety/womenssafety
[147] Meta, 'Our approach to addressing bullying and harassment', *Newsroom,* 9 November 2021, https://about.fb.com/news/2021/11/how-meta-addresses-bullying-harassment

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 (in addition to other Tools already mentioned above, such as Limits):

- **Restrict commenting audience.** We have introduced new tools to give users more control who can comment on their posts on Facebook Feed. Users can control their commenting audience for a public post by choosing from a menu of options. By adjusting the commenting audience, users can further control how they want to invite conversation onto their public posts, and limit potentially unwanted interactions.[148]

- **Hidden Words.** We introduced a tool which will automatically filter messages, comments on posts, and recommended content that contain offensive words, phrases and emojis.[149] We worked with leading anti-discrimination and anti bullying organisations to develop a predefined list of offensive terms that can be filtered; users also have the option to create their own custom lists of words, phrases or emojis that they do not want to see, because we know that different words can be hurtful to different people.

- **New blocking features.** To protect users from unwanted contact, we launched new blocking features so that whenever you decide to block someone on Instagram, users also have the option to block new accounts that person may create.[150] This is designed to help make sure users do not hear from people they have 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 does not want to hear from them.

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 people's relationships offline, not detracting from them.

[148] Meta, 'More control and context in News Feed', *Newsroom*, 21 March 2021, https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed

[149] 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

[150] 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

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

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

Approach to eating disorder content

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

While it can be challenging to proactively identify eating disorder content, due to the many different forms this can take, between January and March 2024, we found and took action on 7.1 million pieces of suicide and self-injury content, including eating disorder content, on Facebook and 5.8 million on Instagram, 99.4% of which was found and actioned before it was reported to us.[151]

Recent tools that we have introduced include nudging teens towards other topics if they have been scrolling on the same topic on Instagram for a while.[152] When someone searches for, or posts, content related to eating disorders or body image issues, they will see a pop-up with tips and an easy way to connect to organisations offering support, including the Butterfly Foundation in Australia.[153] We have also updated this message to make it much more prominent,

[151] Meta, 'Community Standards Enforcement Report - Suicide and Self-Injury', *Transparency Center*, https://transparency.meta.com/reports/community-standards-enforcement/suicide-and-self-injury/facebook/

[152] Instagram, 'New tools and resources for parents and teens in VR and on Instagram', 14 June 2022, https://about.instagram.com/blog/announcements/tools-and-resources-for-parents-and-teens-in-vr-and-on-instagram

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

removing friction by reducing the number of click-throughs to get to helplines and ability to call the helpline within these resources pop-ups.

Approach to suicide and self-injury

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

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

We allow people to discuss topics relating to suicide and self-injury because we want Facebook and Instagram to be spaces where people can share their experiences, raise awareness about these issues and seek support from one another. However, we make content about recovery from suicide of self-harm that is allowed on our services harder for teens to find.[154]

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

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

We have released suicide prevention support on Facebook Live and introduced AI to detect posts that indicate someone may be at risk of imminent harm. And when there is 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.[155]

[154] Meta, 'New protections to give teens more age-appropriate experiences on our apps', *Newsroom*, 9 January 2024, https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps

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

Research

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

According to 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.[156]

Moira Burke, Meta's Data Scientist and Wellbeing Researcher, has undertaken a number of studies on the intersection of wellbeing and social technology.[157] 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.[158]

In a peer-reviewed consensus report led by 12 interdisciplinary experts, the National Academies report concluded: "[t]he committee's review of the literature presented in this chapter and Appendix C did not support the conclusion that social media causes changes in adolescent health at the population level."[159] In fact, the report questions the uniqueness of the crisis, pointing to even higher suicide rates among teens in the early 1990s, before social media. And there's a growing body of research that suggests social media can play a positive role in teens' lives, and provide support to those who may be struggling or are members of marginalised groups. A Pew Research survey also reported that over 90 percent of teens from a nationally representative sample found that social media had a positive or neutral effect on them.[160]

[156] 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

[157] M Burke, *Research,* https://research.facebook.com/?s=burke+moira

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

[159] S Galea, et al. (eds), 'Social media and adolescent health', *National Academies*, p5 http://nap.nationalacademies.org/27396

[160] M Anderson, et al., 'Connection, Creativity and Drama: Teen life on social media in 2022', Pew Research Centre, 16 November 2022, https://www.pewresearch.org/internet/2022/11/16/connection-creativity-and-drama-teen-life-on-social-media-in-2022/

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

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

Additional well-being tools

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

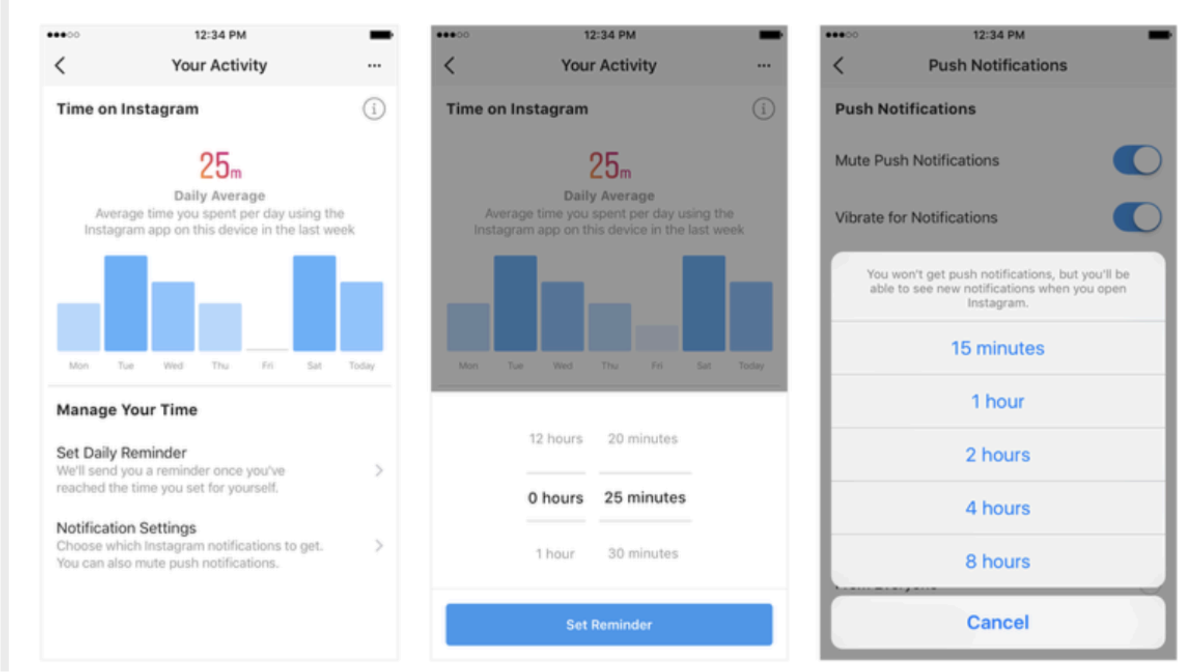
- **Improving Feed quality.** As mentioned above, we have made several changes to Feed to provide more opportunities for meaningful interactions, and reduce passive consumption of low-quality content.[161] We demote things like clickbait headlines and false news. We optimise ranking so posts from the friends you care about most are more likely to appear at the top of your feed. Similarly, our ranking promotes posts that are personally informative. We also redesigned the comments feature to foster better conversations.

- **Activity Dashboard.** The Activity Dashboard, shown in Figure 7 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 have reached the amount of time they want to spend on the app.[162]

Figure 7: Activity Dashboard

[161] M Zuckerberg, Meaningful social interaction post, *Facebook,* 2 November 2017, https://www.facebook.com/zuck/posts/10104146268321841

[162] Meta, ‘New tools to manage your time on Facebook and Instagram’, *Newsroom*, 1 August 2018, https://about.fb.com/news/2018/08/manage-your-time



12:34 PM

Your Activity

Time on Instagram

25m

Daily Average

Average time you spent per day using the Instagram app on this device in the last week

Manage Your Time

Set Daily Reminder

We'll send you a reminder once you've reached the time you set for yourself.

Notification Settings

Choose which Instagram notifications to get. You can also mute push notifications.

12:34 PM

Your Activity

Time on Instagram

25m

Daily Average

Average time you spent per day using the Instagram app on this device in the last week

12 hours 20 minutes

0 hours 25 minutes

1 hour 30 minutes

Set Reminder

12:34 PM

Push Notifications

Push Notifications

Mute Push Notifications

Vibrate for Notifications

You won't get push notifications, but you'll be able to see new notifications when you open Instagram.

15 minutes

1 hour

2 hours

4 hours

8 hours

Cancel

- **Hide Likes on Facebook and Instagram.** We tested hiding like counts to see if it might depressurise people's experience on Instagram.[163] What we heard from people and experts was that not seeing like counts was beneficial for some and annoying to others, particularly because people use like counts to get a sense of what's trending or popular. We now give users the option to hide like counts on all posts they see in their feed. They also have the option to hide like counts on their own posts, so others cannot see how many likes their posts get.

- **Take a Break.** In 2021, we launched a tool called Take a Break on Instagram which empowers people to make informed decisions about how they're spending their time.[164] If someone has been scrolling for a certain amount of time, we ask them to take a break from Instagram and suggest that they set reminders to take more breaks in the future. We also show them expert-backed tips to help them reflect and reset.

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

[163] Meta, 'Giving people more control on Instagram and Facebook', *Newsroom*, 26 May 2021, https://about.fb.com/news/2021/05/giving-people-more-control

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

wellbeing, a Family Centre focused on support for families' online experiences, and an Education Hub with third party resources on conversations about online experiences at home.[165]

[165] See Meta, *Safety Center*, https://facebook.com/safety, Meta, *Family Center,* https://familycenter.meta.com/, and Meta, *Education Hub*, https://familycenter.meta.com/education/