Meta’s Submission to *Select Committee on Adopting Artificial Intelligence Inquiry*

JUNE 2024

Executive summary

Meta welcomes the opportunity to contribute to the inquiry of the Senate Select Committee on Adopting Artificial Intelligence into the opportunities and impacts for Australia arising out of the uptake of artificial intelligence (AI) technologies in Australia (Inquiry).

Since the earliest days of News Feed in 2006, AI has been fundamental to our family of apps - whether it's feed ranking, community discovery, personalised ads, or content moderation. Many of these applications of AI are underpinned by breakthroughs made by our Fundamental AI Research (FAIR) team,[1] which was established over a decade ago.

Generative AI models are enabling entirely new classes of products and experiences, which we are developing to help users better connect and express themselves on our family of apps, as well as better serve businesses that use our platforms to reach and engage with customers. Recently, in April, we announced the release of a new Meta AI assistant, together with our most advanced open large language model, Meta Llama 3, in Australia. Meta AI - which is available on-platform as well as on desktop - can provide recommendations, offer how-to advice, help with writing, create images, and answer questions.[2]

Meta AI is built on Meta Llama 3, our next-generation large language model. Our latest release features pretrained and instruction-fine-tuned language models with 8B and 70B parameters that can support a broad range of use cases. This next generation of Llama demonstrates state-of-the-art performance on a wide range of industry benchmarks and offers new capabilities, including improved reasoning.[3]

We are building AI to drive open innovation and improve experiences across our family of apps and services that are used by billions of people around the world. This presents significant benefits and opportunities for us as a company, our users, and society as a whole. But it also requires new ways of thinking on issues such as privacy, security, fairness and transparency which require broad input.

We are committed to continuing to address hard questions around issues such as privacy, fairness, transparency, and safety. These are questions that no single company can

[1] Meta Research, *Innovating with the freedom to explore, discover and apply AI at scale*, https://ai.meta.com/research/

[2] Meta, 'Meet Your New Assistant: Meta AI, Built With Llama 3', Newsroom, 18 April 2024, https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/

[3] Meta, 'Introducing Meta Llama 3: The most capable openly available LLM to date', *Meta AI blog*, 18 April 2024, https://ai.meta.com/blog/meta-llama-3

answer, which is why we proactively work on them with external stakeholders in multiple different fora. We stress test our products to improve safety performance and regularly work with policymakers, experts in academia and civil society, and others in our industry to advance the beneficial and responsible use of AI.

As the capability and ubiquity of AI has grown, so have calls for greater transparency and control over how the technology is developed and used - from developers and end users to governments and civil society. If people cannot understand how or why an AI model or product behaves the way it does, or feel that they or others have insufficient control over it, it will be difficult to build the trust necessary for the widespread uptake of these technologies.

We have developed and implemented several transparency measures aimed at addressing these concerns and increasing trust in our AI technologies, including on-platform transparency tools for consumers, systems cards and labeling.

We also participate in cross-industry and multi-stakeholder fora - some of which we co-founded - focused on advancing AI responsibly such as the AI Alliance and Partnership on AI. We are also a founding member of ML Commons,[4] an AI engineering consortium, built on a philosophy of open collaboration to improve AI systems.

And finally, we have participated in many of the international AI governance engagements that have taken place in recent months. We supported the G7 Hiroshima AI Process[5], participated in the Bletchley UK AI Summit[6], endorsed the Munich AI Tech Accord on deceptive AI election content[7], and signed the Seoul Frontier AI Safety Commitments[8]. As AI Safety Institutes[9] 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 welcome the opportunity to provide more details about all of these in our submission below.

[4] ML Commons, 'Better AI for Everyone', https://mlcommons.org/
[5] https://www.threads.net/@nickclegg/post/CzCimNDPs5_
[6] https://www.threads.net/@nickclegg/post/CzESxRHLHTA
[7] *A Tech Accord to Combat Deceptive Use of AI in 2024 Elections*, https://www.aielectionsaccord.com/
[8] UK Government, Press release: 'Historic first as companies spanning North America, Asia, Europe and Middle East agree safety commitments on development of AI'. 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
[9] 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).

Table of contents

Using AI to help ensure a safer online environment

We use AI to help ensure a safer online environment for users on our platforms and more broadly. Below are some of the beneficial use cases for AI at Meta.

Combating harmful content and behaviour

Billions of people around the world use Meta's services every day. Hence, detecting and combating harmful content and behaviour at scale is a significant challenge. AI technology provides opportunities to detect harmful content before people need to see it.

While human review continues to play an important role in relation to reviewing certain types of harmful content, AI will be a more effective approach in many instances. For example, AI can moderate content at a scale beyond what humans can achieve, and it also lessens the need for human reviewers in some instances where we want to avoid humans needing to be exposed to the content (for example, in relation to child sexual abuse material).

In the last five or so years, we have had a strong focus on using AI to help enforce our Community Standards,[10] which are the rules that set out what people can or cannot do on Facebook and Instagram. Our ability to use AI to detect and action harmful content proactively has been improving over time.

Our work to combat hate speech online provides an instructive case study. Hate speech is traditionally one of the most challenging types of online content to proactively detect because it is so context-dependent. Five years ago, the volume of hate speech we removed was lower than other categories of harmful content, which meant a high degree of human reporting, review and assessment as needed. When we first started releasing our transparency report in 2017, we removed 1.8 million pieces of hate speech globally, 25 percent of which was detected proactively via AI. Since then, after very significant investments in AI, our proactive detection of hate speech has increased significantly. In Q4, 2023, we removed 7.4 million pieces of hate speech on Facebook and Instagram each, of which 94.5% and 97.3% respectively was detected proactively via AI.[11]

[10] Meta, *Facebook Community Standards,* Transparency Center, https://transparency.fb.com/en-gb/policies/community-standards/

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

We have also significantly cut the prevalence of hate speech content within the last few years (from 0.10 to 0.11 per cent in Q3 2020, down to 0.01-0.02% in Q4, 2023).[12] Prevalence measures the number of views of violating content, divided by the estimated number of total content views on Facebook or Instagram.[13]

We continue to invest in this space, as harmful content continues to evolve - whether through events or by people looking for new ways to evade our systems - and it is crucial for AI systems to evolve alongside it.

This includes working with researchers and experts to try and optimise AI. For example, we have run detection challenges relating to specific types of harmful content like deepfakes[14] and hateful memes.[15]

Our ranking algorithms are also used to reduce the distribution of content that does not violate our Community Standards but is otherwise problematic. This includes clickbait, unoriginal news stories, and posts deemed false by one of the 90 independent fact checking organisations around the world who review content in more than 60 languages. (We outline this in more detail in our discussion of our Content Distribution Guidelines below.)

Promoting age-appropriate experiences online

Protecting our users - particularly young people - is of paramount importance to us in providing our services. Understanding how old someone is underpins these efforts, but it is not an easy task. Finding new and better ways to understand people's ages online is an industry wide challenge. For large-scale companies like Meta, AI is one of the best tools we have to help us tackle these types of challenges at scale.

Over the past decade, in consultation with experts in adolescent development, psychology and mental health, we have developed over 50 tools, features and resources designed to protect young people as well as support them and their parents across our

[12] Meta, *Community Standards Enforcement Report*, Transparency Center, https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/; see also Meta, 'Hate Speech Prevalence Has Dropped by Almost 50% on Facebook', Newsroom, 17 October 2021, https://about.fb.com/news/2021/10/hate-speech-prevalence-dropped-facebook/

[13] Meta, *Prevalence*, https://transparency.fb.com/en-gb/policies/improving/prevalence-metric/

[14] Meta, 'Creating a dataset and a challenge for deepfakes', *Meta AI blog*, 5 September 2019, https://ai.facebook.com/blog/deepfake-detection-challenge/?utm_source=hp

[15] Meta, 'Hateful memes challenge and dataset for research on harmful multimodal content', *Meta AI blog*, 12 May 2020, https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set

apps and technologies.[16] This includes automatically placing all teens into the most restrictive content control settings on Instagram and Facebook and hiding results in Instagram search related to suicide, self-harm and eating disorders.[17]

These controls put a number of default protections in place for those under the age of 16 (or under 18 in certain countries). They also help to empower young people to make the right choices about their experience online, and the information they want to see and share. However, people do not always share their correct age online, and we have seen in practice that misrepresentation of age is a common problem across the industry.

To address this, in June 2022, we shared details about an AI model we have developed to help detect whether someone is a teen or an adult.[18] The job of our adult classifier is to help determine whether someone is an adult (18 and over) or a teen (13–17). The role of our adult classifier is important because, for example, correctly categorising adults is important not only because it allows them to access services and features that are appropriate for them, but also because it helps mitigate risks and child safety issues that could arise on platforms where adults and teens are both present. We do not allow adults to message teens that do not follow them, for example.

Our adult classifier has significantly improved our ability to provide age-appropriate experiences to the people who use our services, but there is room to improve on this work. We are continuously testing new types of signals that might improve our ability to detect whether someone is a teen or adult. Our goal is to expand the use of our AI more widely across Meta technologies and in more countries globally.

[16] See, for example, 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/; Meta, ‘Protecting Teens and Their Privacy on Facebook and Instagram’, Newsroom, 21 November 2022, https://about.fb.com/news/2022/11/protecting-teens-and-their-privacy-on-facebook-and-instagram/; Meta, ‘Giving Teens and Parents More Ways to Manage Their Time on Our Apps’, Newsroom, 27 June 2023, https://about.fb.com/news/2023/06/parental-supervision-and-teen-time-management-on-metas-apps/. For a timeline of some of the many tools and features that we've developed across our apps and technologies, see Meta, ‘Our tools, features and resources to help support teens and parents’, https://www.meta.com/en-gb/help/policies/safety/tools-support-teens-parents/.

[17] 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/

[18] Tech at Meta Blog, ‘How Meta uses AI to better understand people’s ages on our platforms’, 22 June 2022, https://tech.facebook.com/artificial-intelligence/2022/6/adult-classifier/

Using AI for Australia's economic and social benefit

Providing more personalised online experiences

There is a surplus of information and content online. Consequently, it can be a major challenge for individuals to easily find the people, information and experiences that are useful, meaningful and enjoyable for them.

For services like Facebook and Instagram, personalisation is at the heart of the experience. People use our services to connect with family and friends they know, to find communities that they would like to be a part of, and to pursue their interests. We are transparent about how we use AI to make recommendations for people or content that our users may want to engage with.

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. However, as more people started using our services, more content was shared and 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.[19]

We provide this personalised experience via AI. Our ranking algorithms use thousands of signals to rank posts for each person's Feed with this goal in mind.[20] 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 to them, every time they come to Facebook or Instagram.

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.

[19] See, for example, A Mosseri, 'Instagram Ranking Explained', 31 May 2023, https://about.instagram.com/blog/announcements/instagram-ranking-explained/

[20] A Lada, M Wang, '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/

One way we measure whether something creates long-term value for a person is to ask them. For example, we survey people[21] 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.[22] Then we can take each prediction into account for a person based on what people tell us (via surveys) is worth their time.

However, AI does not just bring benefits in terms of convenience, ease or helping people discover new online content; it also brings significant economic benefits.

Many Australian businesses, especially small businesses, benefit from using personalised advertising because it is more efficient and allows them to better reach the right consumer for their business and compete with larger established businesses.

Even just a few years ago, effective advertising was simply not an option for many Australian small businesses: either because it was too expensive (for example, a commercial on free-to-air TV) or too inefficient (for example, newspaper ads which would only be relevant to a subset of a newspaper's readers).

Innovation in advertising (in particular, personalised advertising) has transformed and improved the options available to small businesses for effective advertising.

Firstly, personalised advertising has driven down the cost of advertising overall. According to the Progressive Policy Institute, the share of GDP that is spent on advertising in Australia has dropped 26 per cent from 1991-2000 to 2010-2018. And globally, internet advertising has dropped in price by 42 per cent from 2010 to 2019 (at the same time that other forms of advertising increased in price), due to innovation and advancements in targeting that have made advertising more efficient.[23] These developments are good for advertisers like small businesses and the benefits flow through to consumers, since lower advertising costs means lower prices for the items they buy.

Secondly, it has made advertising much more effective. There is a much greater level of transparency and measurement for advertisers' return on investment when using personalised advertising compared to other forms of advertising.

[21] R Sethuraman, 'Using surveys to make News Feed more personal', Newsroom, 16 May 2019, https://about.fb.com/news/2019/05/more-personalized-experiences/

[22] Meta, How users help shape Facebook, Newsroom, 13 July 2018, https://about.fb.com/news/2018/07/how-users-help-shape-facebook/ ; A Gupta, 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/

[23] M Mandel, *The Declining Price of Advertising: Policy Implications,* https://www.progressivepolicy.org/issues/regulatory-reform/the-declining-price-of-advertising-policy-implications-2/

Personalised advertising has become even more important for Australian small businesses as they recover from the COVID-19 pandemic and associated economic crises. A 2021 report by Deloitte found that 82 per cent of Australian small businesses reported using free, ad-supported Meta apps to help them start their business.[24] It also found that 71 per cent of Australian small businesses that use personalised advertising reported that it is important for the success of their business. Particularly over the past few years, personalised advertising has helped businesses target new customers as they have needed to pivot away from bricks-and-mortar operations during the pandemic, and then pivot back to support the economic recovery.

Consumers also benefit from personalised advertising because they receive advertisements that are more relevant and tailored to their interests. Personalised advertising enables them to discover relevant content (like new brands, new travel destinations or new communities of interest) and find products and services that are more likely to be meaningful and engaging to them.

Further evidence of the benefit of AI-driven advertising is found in research that shows that users prefer personalised advertising to non-targeted advertising: research found that *“the high personalization ad was clearly preferred to the low personalization ad”* by participants in the research, and those users would “*rather share their clicking behaviour and receive behavioural targeted and therefore relevant ads, than random ads”*.[25] The UK Centre for Data Ethics and Innovation described it as: “*[p]eople do not want targeting to be stopped*” and that most people see *“the convenience of online targeting as a desirable feature of using the internet”*.[26]

[24] Deloitte, ‘Dynamic Markets Report: Australia - unlocking small business innovation and growth through the personalised economy’, May 2021, https://scontent-syd2-1.xx.fbcdn.net/v/t39.8562-6/10000000_4303078769743544_7237603050373993547_n.pdf?_nc_cat=109&ccb=1-7&_nc_sid=e280be&_nc_ohc=KTurj8Ra4v4Q7kNvgGANm6Z&_nc_ht=scontent-syd2-1.xx&oh=00_AYAaPwdZxUpcoitpf5iKpbeHduvWdJFVLi59_z7vZsJD9A&oe=6662F549

[25] M Walrave, K Poels, M Antheunis, E Van den Broeck and G van Noort, *Like or Dislike? Adolescents Responses to Personalized Social Network Site Advertising*, Journal of Marketing Communications, Vol. 24, No. 6, 2018, pp. 607, 609, available at: https://www.tandfonline.com/doi/abs/10.1080/13527266.2016.1182938?journalCode=rjmc20; see also, NS Sahni, CS Wheeler, and C Pradeep, ‘Personalization in Email Marketing: The Role of Noninformative Advertising Content,’ Marketing Science, Vol. 37. No. 2, 2018, pp. 241, available at: https://pubsonline.informs.org/doi/10.1287/mksc.2017.1066)

[26] Centre for Data Ethics and Innovation, *Review of online targeting: Final report and recommendations*, February 2020, pp. 6, 48, available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/864167/CDEJ7836-Review-of-Online-Targeting-05022020.pdf.

Supporting innovation

The AI innovations that companies like Meta invest in will, as with many technological innovations, provide additional benefits for users to those they already experience from the use of online services.

As one example, in May 2023, as part of our long-term effort to build language and machine translation (MT) tools that will include most of the world's languages, we announced a series of AI models - our Massively Multilingual Speech (MMS) AI research models - that could make it easier for people to access information and use devices in their preferred language. MMS models expand text-to-speech and speech-to-text technology from around 100 languages to more than 1,100 — more than 10 times as many as before — and can also identify more than 4,000 spoken languages, 40 times more than before. There are also many use cases for speech technology that can be used in a person's preferred language and can understand everyone's voice. We are open-sourcing our models and code so that others in the research community can build on our work and help preserve the world's languages and bring the world closer together.[27]

We can also see the benefits of AI that can be quickly adapted to support public policy goals, such as public health. The use of AI-driven forecasting models during the COVID pandemic provides an example. From April 2020, we created and shared high-quality, localised COVID-19 forecasting models using AI technology to help healthcare providers and emergency responders determine how best to plan and allocate their resources in their particular area. This helped researchers, public health experts, and organisations better understand the spread of COVID-19 given the number of coronavirus cases changed quickly in different communities around the world. We also open-sourced the entire stack of COVID-19 forecasting models so that response teams, governments, and researchers could use them to further help their communities.

Finally, in April this year Meta released Llama 3 – the next generation of our open source large language model.[28] Large language models — natural language processing (NLP) systems with more than 100 billion parameters — have transformed NLP and AI research over the last few years. Trained on a massive and varied volume of text, they show new capabilities to generate creative text, solve basic maths problems, answer reading comprehension questions, and more. At the same time, we announced the roll out of

[27] Meta, 'Preserving the World's Language Diversity Through AI', Newsroom, 22 May 2023, https://about.fb.com/news/2023/05/ai-massively-multilingual-speech-technology/

[28] Meta, 'Meet Your New Assistant: Meta AI, Built With Llama 3', Newsroom, 18 April 2024, https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/

Meta AI to Australia, one of the world's leading AI assistants built with Llama 3 technology, which Australians can use in feed, chats, search and more across Meta's apps to get things done and access real-time information for free.[29]

Meta has put exploratory research, open source, and collaboration with academic and industry partners at the heart of our AI efforts for over a decade. We have seen first-hand how innovation in the open can lead to technologies that benefit more people. Dozens of large language models have already been released and are driving progress by developers and researchers. They are being used by businesses as core ingredients for new generative AI-powered experiences.

Our open-source large language models are already being used by organisations in education, customer service, research and medicine. For example, companies are using Meta Llama to make education content more localised to students, summarise video calls, and provide medical information in low-resource settings.[30] As just one use case, researchers at EPFL's School of Computer and Communication Sciences and Yale School of Medicine used Llama 2 (released in July 2023), to build Meditron, a suite of open-source large multimodal foundation models tailored to the medical field and designed to assist with clinical decision-making and diagnosis. In working closely with humanitarian organizations like the International Committee of the Red Cross, Meditron has the potential to assist healthcare workers in diagnosing and treating patients in underserved areas, and serve emergency scenarios that require fast and accurate medical response. Following our recent release of Meta Llama 3, the team fine-tuned the new model within 24 hours to deliver Llama-3[8B]-MeditronV1.0,[31] which achieves strong results on leading industry benchmarks in the field, such as question-answering of biomedical exams.[32] The goal is that open models like Meditron can help create equitable access to medical knowledge.

As another example, Netsafe, New Zealand's independent, non-profit online safety charity, uses Meta Llama models to develop a robust redaction tool for harmful digital communications. The tool uses a tuned LLM model and the AI4Privacy dataset to automatically redact sensitive information, enabling faster and more effective harm resolution. Netsafe's data pipeline sources, redacts, and aggregates data from various

[29] Meta, 'Meet Your New Assistant: Meta AI, Built With Llama 3', Newsroom, 18 April 2024, https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/

[30] Meta, 'How Companies Are Using Meta Llama', Newsroom, 7 May 2024, https://about.fb.com/news/2024/05/how-companies-are-using-meta-llama/

[31] https://meditron-ddx.github.io/llama3-meditron.github.io/

[32] Meta, 'How Companies Are Using Meta Llama', Newsroom, 7 May 2024, https://about.fb.com/news/2024/05/how-companies-are-using-meta-llama/

channels, minimising victim impact and augmenting Digital Harm Resolution Officers' skills.[33]

Meta's approach to responsible innovation

At Meta, we have five pillars of responsible AI that inform our work – privacy, fairness and inclusion, robustness and safety, transparency and control and accountability and governance.

Privacy

AI models and tools will continue to be addressed in Meta's robust privacy review process with privacy risks identified, mitigated, evidenced and monitored.

Protecting people's information and giving them control over their data is a company-wide priority. For example, our privacy review process[34] is designed to assess privacy risks that collecting, using, or sharing people's information may present, and to help determine whether steps should be taken to mitigate any identified privacy risks, including through the development and use of AI models and tools.

We have taken several steps to mitigate privacy risks arising from the training and use of our AI models and tools, including:

- Taking steps to remove data from certain sites known to contain a high volume of personal information about individuals.
- We train and fine-tune our generative AI models to limit the possibility of personal information that users may share with our generative AI products (e.g., our Meta AI assistant) from appearing in responses to other people.
- We use a combination of human reviewers and automated technology to review model outputs so we can reduce the likelihood of outputs including personal information (as well as improve product performance).

Safety & security

We are committed to ensuring that our AI is robust and safe for everyone to use. Like any emerging technology, AI must be shown to be robust and safe for it to be trusted and widely adopted.

[33] Meta, 'Community stories: Discover the possibilities of building on Meta Llama', https://llama.meta.com/community-stories
[34] Meta, 'Privacy progress update', https://about.meta.com/uk/privacy-progress/

We favour an open source approach to AI development. By making our AI models openly available to the AI community, we are able to leverage the expertise of many other experts, who are able to inspect, evaluate, and improve our models. This approach is complemented by other initiatives, such as our bug bounty programme that offers rewards for those who identify and report vulnerabilities in our models (within the terms of the programme).

Openly releasing models does not, however, mean that we leave safety issues to the AI community to resolve alone. Before releasing a model, we work to identify, evaluate, and mitigate potential risks through several measures, including pre-deployment risk assessments, red-teaming, and safety fine-tuning.[35] These processes help ensure that our decision to release a model is the right and responsible thing to do.

When we decide to release a model, we empower the AI community to develop and use our models responsibly through resources such as our Responsible Use Guide,[36] which outlines best practices and considerations for developers , together with some mitigation strategies and resources available to developers to address risks at various points in the system. These resources include the evaluation and safety tools that we make openly available as part of our Purple Llama project.[37]

We recognise, however, that not all parts of the AI community intend to use our AI responsibly. We take steps during red-teaming and fine-tuning to reduce the likelihood of our models being misused, or producing unhelpful or unsafe outcomes, including techniques like prompt filtering and blocklist, which are designed to detect and block harmful model prompts and outputs.

Fairness & Inclusion

We develop and scale tests and tools that aim to minimise potential bias and enable more inclusive and accessible AI. One aspect of fairness in AI relates to how AI systems may affect people in diverse groups, or groups that have been historically marginalised in society. We know it is possible for AI systems to learn societal biases through their training data, and the trends in that data or gaps in representation can manifest in a range of potential issues, like producing content that reinforces negative stereotypes or in erasure of identities. At the same time, we understand how critical it is to respect privacy for sensitive demographic data, even when it is being used to measure and close potential fairness gaps.

[35] Meta, ‘Our responsible approach to Meta AI and Meta Llama 3’, 18 April 2024, *Meta AI blog*, https://ai.meta.com/blog/meta-llama-3-meta-ai-responsibility/
[36] Meta, ‘Meta Llama: Responsible Use Guide’, https://llama.meta.com/responsible-use-guide/
[37] See Meta Llama, Making safety tools accessible to everyone, https://llama.meta.com/trust-and-safety/

We are continuing our work to create and distribute more diverse datasets that respect privacy and represent a wide range of people and experience, to enable researchers to better evaluate the fairness and robustness of certain types of AI mode. For example, we have publicly released Casual Conversations v2,[38] a consent-based dataset for evaluating trained models in computer vision and audio applications by measuring their accuracy across a diverse set of ages, genders, languages/dialects, physical attributes, voice timbres, skin tones, and more.

We are also furthering more inclusive and accessible AI by increasing language accessibility and coverage through No Language Left Behind (NLLB),[39] a first-of-its-kind project that open sources, under a non-commercial licence, AI models capable of delivering high-quality translations between 200 languages, including low-resource languages such as Asturian, Luganda, and Urdu. The project aims to give people the opportunity to access and share web content in their native language, and communicate with anyone, anywhere, regardless of their language preferences.

As a real-world example, the technology behind NLLB is supporting Wikipedia editors as they translate information from their native and preferred languages. Wikipedia editors are using the technology to more efficiently translate and edit articles originating in low-resource languages. This helps make more knowledge available in more languages for Wikipedia around the world.

As well as No Language Left Behind, we're also increasing language accessibility and coverage through Massively Multilingual Speech (MMS) and SeamlessM4T. MMS is an AI research model capable of identifying more than 4,000 spoken languages - 40 times more than any known previous technology.[40] The identification capability has led to an expansion of text-to-speech and speech-to-text technologies from around 100 languages to more than 1,100 languages.

SeamlessM4T is the first all-in-one, multimodal, multilingual AI translation and transcription model.[41] This single model can perform speech-to-text, speech-to-speech, text-to-speech, and text-to-text translations for up to 100 languages. The launch of SeamlessM4T also included the release of SeamlessAlign, an open, multimodal

[38] Meta, 'Introducing Casual Conversations v2: A more inclusive dataset to measure fairness', *Meta AI blog*, 9 March 2023, https://ai.meta.com/blog/casual-conversations-v2-dataset-measure-fairness/

[39] Meta, No Language Left Behind: Driving inclusion through the power of AI translation, https://ai.meta.com/research/no-language-left-behind/

[40] Meta, 'Preserving the World's Language Diversity Through AI', Newsroom, https://about.fb.com/news/2023/05/ai-massively-multilingual-speech-technology/

[41] Meta, Bringing the world closer together with a foundational multimodal model for speech translation, *Meta AI blog*, 22 August 2023, https://ai.meta.com/blog/seamless-m4t/

translation dataset including 470,000 hours of speech and text alignments, and SONAR, a suite of speech and text sentence encoders that allow developers to further mine their monolingual datasets.

Transparency & control

We prioritise providing greater transparency to users through measures that explain how our AI-powered products work and enable users to understand when they are engaging with AI-generated content. 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.

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.[42] 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.[43]
- *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.[44]

[42] 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

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

[44] Facebook, '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

- *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.[45]
- *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.[46]
- *Content recommendation controls* - our content recommendation controls - known as "Sensitive Content Control" on Instagram and "Reduce" on Facebook – make it more difficult for people to come across potentially sensitive content or accounts in places like Search and Explore.[47]

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 20 AI System Cards that explain how the AI systems in our products work.[48] They give information, for example, about how our AI systems rank content, some of the predictions each system makes to determine what content might be most relevant to users, as well as the controls users can use to help customise their experience.

This is complemented by Meta's Transparency Center[49] and Privacy 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] The Privacy Center also includes guides about Generative AI and a Teen Generative AI Guide.[52] We also provide a deeper look at the types of signals and prediction models that we use in our ranking systems to reduce

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

[46] Facebook, '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

[47] 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

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

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

[50] Meta, Privacy Center, https://www.facebook.com/privacy/center

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

[52] See information about GenA: 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

problematic content.[53] 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.[54]

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,[55] 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,[56] and Meta support and resources for teens relating to generative AI.[57]

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.

This year, we have begun labeling a wider range of video, audio and image content as "Made with AI" when we detect industry standard AI image indicators or when people disclose that they're uploading AI-generated content.[58] 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.[59] 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.

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

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

[55] Meta, 'How Meta uses information for generative AI models and features', Privacy Center, https://www.facebook.com/privacy/genai

[56] Meta, 'Generative AI at Meta', Privacy Center, https://www.facebook.com/privacy/guide/generative-ai/

[57] Meta, 'Access support and resources for teens', Privacy Center, https://www.facebook.com/privacy/guide/teens/

[58] 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/

[59] 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

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.[60] Continued research and collaboration with experts can assist in supporting technical work that enables AI to be more explainable and predictable.

Governance Frameworks & Partnerships

Responsible AI fora

As noted above, no one single company can address the hard questions around issues such as privacy, fairness, transparency, and safety that arise in relation to AI. This is why we proactively work on them with external stakeholders in multiple different fora and are actively and constructively engaged in global conversations around appropriate governance frameworks.

For example:

- Together with IBM, we established the AI Alliance,[61] a community of over 50 technology creators, developers and adopters collaborating to advance safe, responsible AI rooted in open innovation. The Alliance seeks to (i) build and support open technologies across software, models and tools, (ii) enable developers and scientists to understand, experiment, and adopt open technologies, and (iii) advocate for open innovation with organisational and societal leaders, policy and regulatory bodies, and the public.
    - Members and collaborators comprise companies, academic, and non-profit organisations that favour an open innovation approach to AI.[62] Notably, peer companies that advocate for a more proprietary approach to AI development - e.g., Google, Anthropic, OpenAI, and Microsoft - have not yet joined the Alliance.

[60] Meta Research, '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/
[61] AI Alliance, Building the Open Future of AI, https://thealliance.ai/
[62] See AI Alliance, 'Meet the AI Alliance', https://thealliance.ai/aia-members

- We are a founding member and funding support of the Partnership on AI (PAI),[63] a non-profit community of academic, civil society, industry, and media organisations creating solutions so that AI advances positive outcomes for people and society. We have representation on PAI's board, together with representatives from organisations such as Google Deepmind, Apple, Intel Labs, Microsoft, and OpenAI. We have been involved in several of PAI's initiatives, including:

  - Synthetic Media Framework.[64] A framework that provides specific recommendations on how to develop, create, and share content that has been generated or modified (usually via AI) in a responsible way. Recommendations focus on transparency, disclosure, and provenance measures, and some are tailored for different categories of stakeholder (e.g., those building the technology and infrastructure for synthetic media, those creating it, and those distributing and publishing it).

  - Guidance for Safe Foundation Model Deployment.[65] A framework that provides foundational model providers with a set of recommended practices to follow throughout the deployment process, tailored to the capabilities of their specific model (e.g., narrow or general purpose) and how it is being released (e.g., open source or behind an API).

- We are a founding member of ML Commons,[66] an AI engineering consortium, built on a philosophy of open collaboration to improve AI systems. Its mission is to create AI innovation and increase its positive impact on society by bringing together industry, academia, and non-profit organisations from around the world to develop open industry-standard benchmarks[67] and large-scale, diverse datasets[68] for model evaluation. We have representation on ML Commons' board, together with representatives from organisations including Google, Nvidia, Intel, Qualcomm, and Alibaba.

[63] Partnership on AI, https://partnershiponai.org/
[64] Partnership on AI, 'PAI's Responsible Practices for Synthetic Media A Framework for Collective Action', https://syntheticmedia.partnershiponai.org/
[65] Partnership on AI, 'PAI's Guidance for Safe Foundation Model Deployment
A Framework for Collective Action', https://partnershiponai.org/modeldeployment/
[66] ML Commons, 'Better AI for Everyone', https://mlcommons.org/
[67] ML Commons, 'Benchmarks', https://mlcommons.org/benchmarks/
[68] ML Commons, 'Datasets', https://mlcommons.org/datasets/

- We recently joined the Frontier Model Forum, a non-profit organization and the only industry-supported body dedicated to advancing the safety of frontier AI models.[69]

There is also 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*,[70] 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[71], 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.[72]

We have participated in many of the international AI governance engagements, specifically, we supported the G7 Hiroshima AI Process,[73] participated in the Bletchley UK AI Summit[74] and signed the Seoul Frontier AI Safety Commitments[75]. As AI Safety Institutes[76] 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.[77]

In February 2024, we also signed on to *A Tech Accord to Combat Deceptive Use of AI in 2024 Elections*, pledging to work with other technology companies including Amazon, Google, and OpenAI, to help prevent deceptive AI content from interfering with this

[69] See https://www.frontiermodelforum.org/updates/amazon-and-meta-join-the-frontier-model-forum-to-promote-ai-safety/

[70] US National Archives Federal Register, 'Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', 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

[71] 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

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

[73] https://www.threads.net/@nickclegg/post/CzCimNDPs5

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

[75] 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

[76] 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).

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

year's global elections.[78] This work is bigger than any one company and will require a huge effort across industry, government, and civil society.

We trust that these insights are helpful to the Committee as it undertakes this Inquiry and we welcome the opportunity to continue to collaborate with governments on delivering the many benefits of AI in Australia, whilst working to mitigate risks and addressing policy concerns.

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