Meta's Submission on *Safe & Responsible AI in Australia* Discussion Paper

AUGUST 2023

Executive summary

Meta welcomes the opportunity to contribute to the Australian Department of Industry, Science and Resources' consultation in response to the issues paper *Safe and Responsible AI in Australia*.

Given recent breakthroughs in artificial intelligence (AI) , and generative AI in particular, have captured the public's imagination and demonstrated the potential to help people do incredible things, create a new era of economic and social opportunities, and give individuals, creators, and businesses new ways to express themselves and connect with people – it is timely for a review of Australia's regulatory and governance responses in relation to AI.

Before discussing governance options in relation to AI, to assist the Department with its consultation, we wanted to first share some background about how Meta uses AI and our approach focused on transparency, openness and responsible innovation.

At Meta, we believe AI should benefit everyone – not just a handful of companies. AI innovation is inevitable and AI should be built to benefit the whole of society. Like all foundational technologies – from radio transmitters to internet operating systems – there will be a multitude of uses for AI models, some predictable and some not. And like every technology, AI will be used for both good and misused by good and bad people. While we can't eliminate the risks, we can mitigate them.

As part of considering how to mitigate risks from AI, it is also important to recognise the extent to which AI is already widely deployed within industry especially as part of any discussions about what, if any, new regulatory frameworks may be needed. A review of the existing uses of AI by the industry will assist in identifying the role of AI to comply with regulatory obligations and community expectations in Australia and what adjustments to existing laws may be necessary to address public policy concerns identified from the broader training and deployment of AI systems.

Since the earliest days of Feed in 2006, Meta has used machine learning and AI to power all of our apps and services - whether it's personalised content feeds, keeping our platforms safe, or showing relevant ads. Use of AI on Meta's services has already been generating significant benefits in Australia for some time.

With respect to our ranking, recommendation and discovery engines, AI has personalised the online experience of the millions of Australians who use our services, connecting them with the people, communities and information that is most useful to them. For example, AI-recommended content from accounts you don't follow is now the fastest growing category of content on Facebook's Feed. Reels plays exceed 200 billion per day across Facebook and Instagram. In addition, the economic benefits from the use of AI on Meta's services are also significant –, for example, Australian small and medium businesses having access to efficient and effective targeted advertising. A recent report found that 75% of Australian small to medium enterprises (SMEs) report that Meta technologies enabled their business to market and sell its products and services and 67% of SMEs believe their business is stronger today because of Meta technologies and apps.[1]

AI is also central to our integrity systems, which are designed to protect our platform and our users. Meta's use of AI has, essentially, been supporting many of the Australian government's policy objectives with respect to online safety, the protection of youth well-being, and privacy. For example, AI is proactively detecting more than 90 per cent of the harmful content on Facebook and Instagram that we action for violating our Community Standards[2] (higher in some areas). Meta is also increasingly, for example, using AI to provide more age-appropriate experiences on our services.[3]

To promote greater understanding of the use of AI across our products and integrity systems, Meta has for many years invested in significant and industry leading transparency measures. With respect to content and ads ranking, we have in-product transparency tools[4] and explanations about the policies and principles that guide ranking and recommendations algorithms in our Transparency Center and Help Center.[5]

We also recently released 22 system cards for Facebook and Instagram.[6] They give information about how our AI systems rank content, some of the predictions each system makes to determine what content might be most relevant to you, as well as the controls you can use to help customise your experience.

[1] Thoughtlab, *The Digital Journey of SMEs in Australia*, May 2023, https://thoughtlabgroup.com/the-digital-journey-of-smes-in-australia/

[2] See *Community Standards Enforcement Report,* https://transparency.fb.com/data/community-standards-enforcement/

[3] Tech at Meta Blog, *How Meta uses AI to better understand people's ages on our platforms*, 31 March 2021, https://tech.facebook.com/artificial-intelligence/2022/6/adult-classifier/ (June 2022)

[4] *See e.g.,* Meta Newsroom, *'More control and context in News Feed'*, 31 March 2021, https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/; Meta, *'Understand why you're seeing certain ads and how you can adjust your ad experience',* 11 July 2019, https://about.fb.com/news/2019/07/understand-why-youre-seeing-ads/

[5] *See e.g.,* Facebook Help Center, *'What are recommendations on Facebook?'* https://www.facebook.com/help/1257205004624246/ ; Instagram Help Center, *'Recommendations on Instagram'* https://help.instagram.com/313829416281232/?helpref=uf_share

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

With respect to the AI used as part of our integrity systems, we publish a quarterly Community Standards Enforcement Report[7] that outlines how much content we are actioning for violations of our policies and, importantly, how much is identified proactively by our technology before anyone reports it to us.

In addition to the use of "Classic AI" across our product and integrity systems, Meta has been investing in new generative foundation models that are enabling entirely new classes of products and experiences ("Generative AI").[8] Innovations driven by this technology will provide enormous benefits for people and society. We have released over 1,000 models and AI databases on non-commercial licences for researchers, so that they can benefit from the computing power we are able to deploy and pursue their own research openly and safely. For example, we are pursuing language translation projects[9] which will unlock access to the internet for the billions of people around the world not currently able to view online content in their preferred language.

Most recently, we have been releasing – free for research and now commercial use – large language models (Llama). For our Llama 1 model, there were over 100,000 requests for access from individuals and organisations in the research community. This past month, we released Llama 2 – our latest large language model, ranging from 7 billion to 70 billion parameters – and are making it available free of charge for research and commercial use.[10] We can envisage use cases such as credit card companies using it to improve anomaly detection and fraud analysis, medical professionals making more accurate diagnoses by identifying anomalies in medical images, and businesses using it for organisational tasks.

By democratising access, via this open approach, to foundation language models, potential toxicity, bias, bugs and vulnerabilities can be continuously identified and mitigated in a transparent way by an open community. Advancing our efforts towards an open approach for AI has been welcomed by more than 90 global academics, policy makers and technology companies.[11]

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

[8] Classic AI is known for being able to analyse large amounts of data which can be used, for example, to classify and label content (e.g., integrity models), or predict what content users will find most relevant or valuable (e.g., ranking and recommender models). Generative AI is differentiated through its ability to create new content using existing text, audio, images, or videos.

[9] Meta Newsroom, 'Inside the Lab: Building for the Metaverse with AI', *Meta Newsroom*, 23 February 2022, https://about.fb.com/news/2022/02/inside-the-lab-building-for-the-metaverse-with-ai/.

[10] Meta AI, *Introducing Llama 2,* http://ai.meta.com/llama

[11] Meta Newsroom, *Statement of Support for Meta's Open Approach to Today's AI,* June 2023 https://about.fb.com/news/2023/07/llama-2-statement-of-support/

We are also conscious of the importance of protecting fairness, inclusion and privacy in the context of AI. We have released a set of Responsible Innovation Principles (including five pillars of responsible AI) to guide our work, and have launched a product called Fairness Flow to help internal teams analyse how AI works, so that they can spot risks or unintended consequences. Additionally, our open source approach to large language models and accompanying support resources are designed to support the responsible build of AI. And finally, we regularly publish our research and updates about our AI work to provide transparency and deepen industry expertise.[12]

Given the amount of progress occurring within industry, it is an opportune time for the Australian Government to engage in conversations about the right regulatory frameworks for emerging technology such as AI. Meta has been at the global forefront of calling for updated regulation for the internet.[13]

Against this background and with respect to the specific questions included in the consultation paper, we encourage the Department to consider the following principles for AI regulation:

- Use definitions that strike the right balance between precision and flexibility and consistent with international definitions such as that adopted by the OECD Expert Group on AI
- Review existing regulatory frameworks that govern many of the policy concerns raised in the context of AI to assess the extent to which they are fit-for-purpose already or may need to be adjusted
- Consider how AI regulation can be built upon existing legislation that already impacts AI, without creating tension with existing obligations
- Adopt a framework, when assessing Generative AI research models, that breaks out the policy issues that this new technology may present into three areas – research model training data, evaluation of user inputs and model outputs – to allow proportionate identification of potential policy responses
- Ensure AI regulation is principle-based and adopts a pro-innovation, risk-based approach, focused on the uses of the technology and not the technology specifically

[12] See e.g., Meta AI, *Research,* https://ai.meta.com/research/
[13] 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; and M Zuckerberg, 'Big tech needs more regulation', *Financial Times*, 16 February 2020, https://www.ft.com/content/602ec7ec-4f18-11ea-95a0-43d18ec715f5.

- Encourage open innovation and competition so that AI benefits everyone–not just a handful of companies—and is built by an AI research community to benefit the whole-of-society
- Design any AI regulation as a product of collaboration amongst multiple stakeholders and benchmarked against many of Australia’s regional allies such as Japan, the US and Singapore

We are still at the very early stages of AI technology, and there is an exciting opportunity for the global community working on AI to strive towards the innovations that will help to solve our greatest challenges. With respect to regulation for the frontier of this innovation, the key focus must be to develop regulations that are broad and flexible enough to adapt to future technologies while not overly restrictive to the point of suppressing valuable and beneficial innovations in, and uses of, AI technology. For this reason, we need collaborative policymaking to ensure an appropriately balanced approach. Many AI systems at issue are profoundly complex, often the result of the combined effort of thousands of engineers around the world, and there are unanswered questions about how to provide meaningful transparency while preserving privacy and trade secrets all while maintaining a fertile ground for innovation.

We welcome the opportunity to provide more details about all of these in our submission below. We also look forward to working with the Australian Government on how we can progress these public policy objectives together.

Table of contents

Meta's work on artificial intelligence and algorithms

Meta uses AI in a wide variety of ways as part of our content governance and integrity systems, to optimise ads and drive sales for small businesses, and to support innovation, including in the use of large language models for socially useful purposes. Below is an overview of some of these 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 combatting 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,[14] 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.

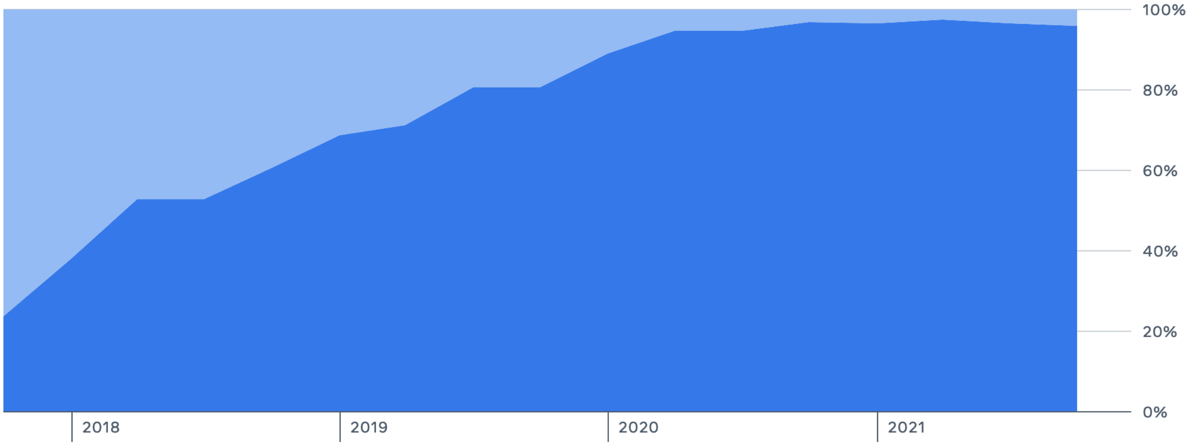
Our work to combat hate speech online provides an instructive case study. Hate speech is traditionally one of the most challenging types of online content to proactively detect because it is so context-dependent. Five years ago, the volume of hate speech we removed was lower than other categories of 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, our proactive detection of hate speech has increased significantly. By the end of 2020, we removed 26.9 million pieces of hate speech, 95 percent of which was detected proactively via AI.

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

We have also significantly cut the prevalence of hate speech content by more than half within the last year alone (from 0.07 to 0.08 per cent in Q3 2020, down to less than 0.03 per cent in Q3 2021). Prevalence measures the number of views of violating content, divided by the estimated number of total content views on Facebook or Instagram.[15]

The improvement in the detection rate is shown in the following **Figure 1**. These advancements were made after very significant investments in AI.

Figure 1: Proactive detection rate for hate speech on Facebook
(ie. Of the violating content we detected for hate speech, how much did we find before people reported it?)



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's crucial for AI systems to evolve alongside it.

It typically takes several months to collect and label thousands, if not millions, of examples necessary to train each individual AI system to spot a new type of content.[16] To tackle this, we built and deployed Few-Shot Learner (FSL), an AI technology that can adapt to take action on new or evolving types of harmful content within weeks instead of months. This new AI system uses a method called "few-shot learning," in which models start with a general understanding of many different topics and then use much fewer — or sometimes zero — labelled examples to learn new tasks. FSL can be used in more than 100 languages and learns from different kinds of data, such as images and text. This new technology will help augment our existing methods of addressing harmful content.

[15] Meta, *Prevalence*, https://transparency.fb.com/en-gb/policies/improving/prevalence-metric/.
[16] Meta, 'Our new AI system to help tackle harmful content', *Meta Newsroom*, 8 December 2021, https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/.

There are a range of other technologies that can help us identify harmful content faster, across languages and content type (i.e. text, image, etc.) such as RIO,[17] WPIE[18] and XLM-R[19]. Last year we announced the first-ever high-performance self-supervised algorithm for speech, vision and text, which will enable much more adaptable AI that will be able to perform tasks in the future well beyond what is possible today.[20]

We also work 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[21] and hateful memes.[22]

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 more than 80 independent fact checking organisations that evaluate Facebook content. (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.

[17] 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/

[18] 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/

[19] 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/

[20] Meta, 'Introducing the first self-supervised algorithm for speech, vision and text', *Meta Newsroom*, 20 January 2022, https://about.fb.com/news/2022/01/first-self-supervised-algorithm-for-speech-vision-text/.

[21] Meta AI, '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.

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

Facebook and Instagram already have a number of measures in place to provide an age-appropriate experience to those between the ages of 13 and 18, including but not limited to:[23]

- **Defaulting new teen accounts to private.** We default all new Instagram users who are under the age of 16 in Australia onto a private account.
- **Implementing privacy-protective default settings.** There are a range of other default limits that are placed on a minor's account on Facebook. For example, profiles of minors cannot be found on Facebook nor do we allow search engines to index profiles of minors off our platform; Post and Story audiences are defaulted to Friends (rather than public); and Location is turned off by default.
- **Encouraging existing teen accounts to be private.** For young people who already have a public account on Instagram, we show them a notification highlighting the benefits of a private account and how to change their privacy settings. We will still give young people the choice to switch to a private account or keep their current account public if they wish.

These controls put a number of default protections in place for those under the age of 18. They also help to empower young people to make the right choices about their experience online, and the information they want to see and share. 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. 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 don't allow adults to message teens that don't follow them, for example.

To develop our adult classifier, we first train an AI model on signals such as profile information, like when a person's account was created and interactions with other profiles and content. To evaluate the performance of the model, we develop an "evaluation dataset." That dataset is created by having teams manually review certain data points that we believe to be strong signals of age, such as birthday posts. Identifying details are removed before these posts are shared with the team to make a determination about the age of the person who posted it. Once the team has made that determination, they label

[23] Meta, 'Giving young people a safer, more private experience on Instagram', *Meta Newsroom,* 27 July 2021, https://about.fb.com/news/2021/07/instagram-safe-and-private-for-young-people/

the data with a note indicating whether the post was made by an adult or a teen. These labelled data points then make up our evaluation dataset.

We then evaluate our classifier on a country-by-country basis. Before applying the classifier to a new country, we look at its performance across several criteria, including overall accuracy and accuracy across different groups of people. For example, since we use interactions with content as a signal, we look at how our model performs for people who have not been on our platform for very long and therefore have not yet interacted with much content. But the work is not done once the classifier is up and running. To check that our determinations are up-to-date, we regularly rerun the classifier to include the latest information.

Each time we retrain the model, we check its age detections against the labelled evaluation dataset to measure the model's accuracy. We have a sophisticated framework to ensure that our evaluation dataset is representative of the people using our services and that our model accuracy metrics are generalizable to the population of people using our services. We also recognise that no matter how accurate an AI system is, it can occasionally get calculations wrong. When that happens, we give people options to manually verify their age.[24]

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.

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.

[24] Meta Newsroom, *Introducing news ways to verify age on Instagram, June 2022,* https://about.fb.com/news/2022/06/new-ways-to-verify-age-on-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. 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.[25]

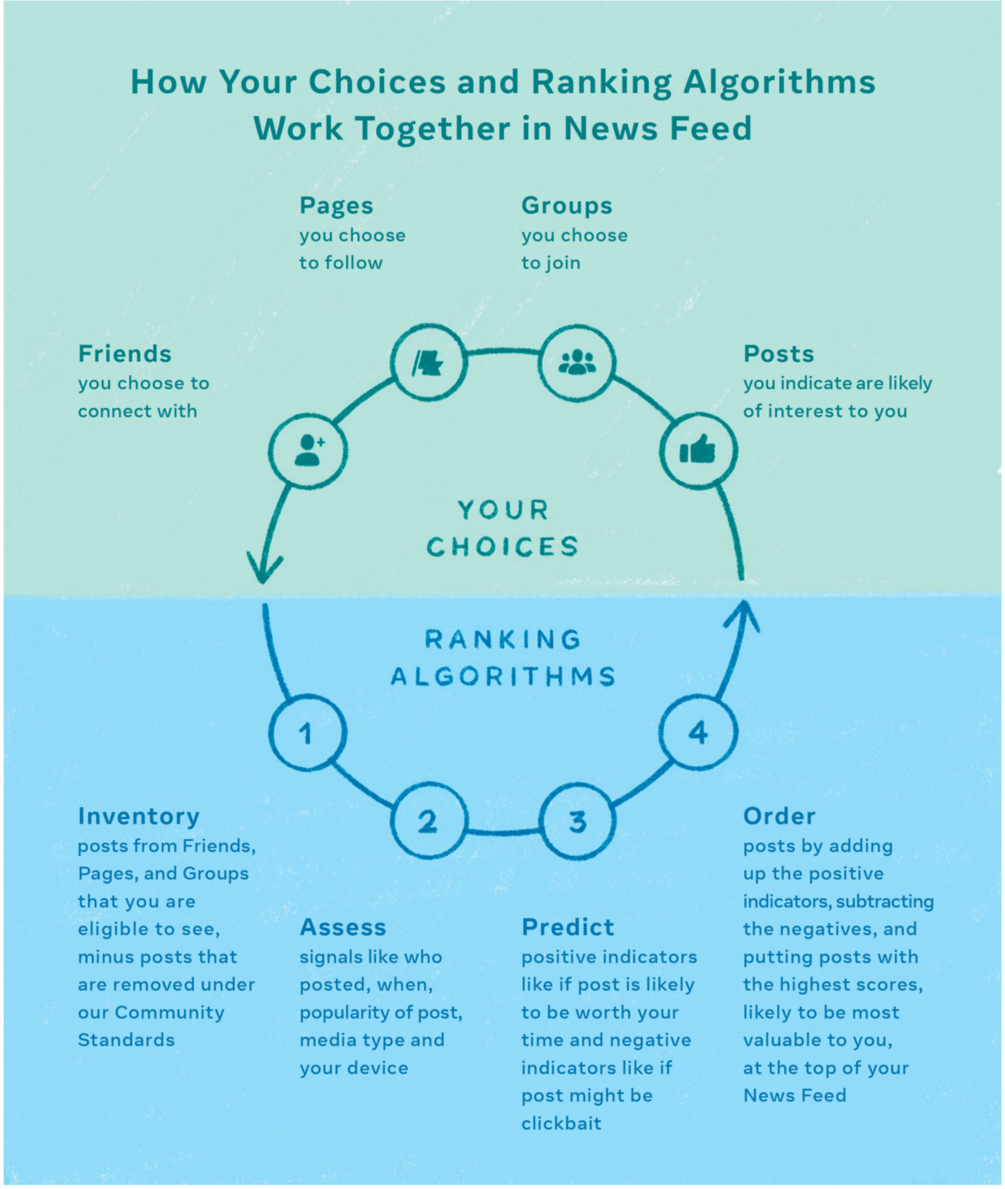
We provide this personalised experience via AI. Our ranking algorithms use thousands of signals to rank posts for each person's Feed with this goal in mind.[26] As a result, each person's Feed is highly personalised and specific to them. Our ranking system personalises the content for over a billion people and aims to show each of them content we hope is most valuable to them, every time they come to Facebook or Instagram.

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

From that initial inventory, thousands of signals are assessed for these posts, like who posted it, when, whether it's a photo, video or link, how popular it is on the platform, or the type of device you are using, see Figure 2 below for more detail. In the next step from there, our ranking algorithms use these signals to predict how likely the post is to be relevant and meaningful to a person: for example, how likely a person might be to "like" it or find that viewing it was worth their time. The goal is to make sure people see what they will find most meaningful — not to keep people glued to their smartphone for hours on end.

[25] *See e.g.,* A Mosseri, 'Shedding more light on how Instagram works', *Instagram Blog,* 8 June 2021, https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works
[26] A Lada, M Wang, 'How does News Feed predict what you want to see?', *Meta Newsroom,* 26 January 2021, https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/

Figure 2: How your choices and ranking algorithms work together


How Your Choices and Ranking Algorithms Work Together in News Feed
Pages
you choose to follow
Groups
you choose to join
Friends
you choose to connect with
Posts
you indicate are likely of interest to you
YOUR CHOICES
RANKING ALGORITHMS
1
2
3
4
Inventory
posts from Friends, Pages, and Groups that you are eligible to see, minus posts that are removed under our Community Standards
Assess
signals like who posted, when, popularity of post, media type and your device
Predict
positive indicators like if post is likely to be worth your time and negative indicators like if post might be clickbait
Order
posts by adding up the positive indicators, subtracting the negatives, and putting posts with the highest scores, likely to be most valuable to you, at the top of your News Feed

One way we measure whether something creates long-term value for a person is to ask them. For example, we survey people[27] to ask how meaningful they found an interaction or whether a post was worth their time, so that our system reflects what people enjoy

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

and find meaningful.[28] Then we can take each prediction into account for a person based on what people tell us (via surveys) is worth their time.

While a post's engagement — or how often people like it, comment on it, or share it — can be a helpful indicator that it's interesting to people, this survey-driven approach, which largely occurs outside the immediate reaction to a post, gives a more complete picture of the types of posts people find most valuable, and what kind of content detracts from their Feed experience.

We also use AI to make recommendations for people or content that our users may want to engage with. Similar to ranking, we prioritise content that is "unconnected" (ie. a person doesn't already follow that Page or Instagram account and isn't connected with that person) by looking at signals like what posts you've liked, saved, and commented on in the past. Once we've found a group of photos and videos a person might be interested in, we then order them by how interested we think a person might be in each one.[29]

However, AI doesn't 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

[28] Meta, How users help shape Facebook, *Meta 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, *Meta Newsroom,* 22 April 2021, https://about.fb.com/news/2021/04/incorporating-more-feedback-into-news-feed-ranking/

[29] A Mosseri, 'Shedding more light on how Instagram works', *Instagram Blog,* 8 June 2021, https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works

the same time that other forms of advertising increased in price), due to innovation and advancements in targeting that have made advertising more efficient.[30] 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.

Personalised advertising has become even more important for Australian small businesses as they recover from the COVID-19 pandemic and associated economic crisis. 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.[31] 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 two 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"*.[32] The UK Centre for Data Ethics and Innovation described it as: "*[p]eople do not want targeting to*

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

[31] Deloitte, 'Dynamic Markets Report: Australia - unlocking small business innovation and growth through the personalised economy', Meta Australia blog, October 2021, https://australia.fb.com/economic-empowerment/

[32] 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)

be stopped" and that most people see *"the convenience of online targeting as a desirable feature of using the internet"*.[33]

Supporting innovation

The AI innovations that companies like Meta invest in will, as with many technological innovations, provide exciting additional benefits for users. Take for example, the diversity and inclusion benefits that will result from our work on projects relating to language translation.[34] Nearly half the world's population - billions of people - are not able to access online content in their preferred language. Today's machine translation systems are improving rapidly, but they still rely heavily on learning from large amounts of textual data, so they do not generally work well for low-resource languages, i.e., languages that lack training data, and for languages that don't have a standardised writing system. This problem will be known acutely by First Nations people in Australia and the immediate region.

Meta has announced a long-term effort to build language and machine translation (MT) tools that will include most of the world's languages. This includes two new projects.

1. *No Language Left Behind*, where we are building a new advanced AI model that can learn from languages with fewer examples to train from, and we will use it to enable expert-quality translations in hundreds of languages, ranging from Asturian to Luganda to Urdu.
2. *Universal Speech Translator*, where we are designing novel approaches to translating from speech in one language to another in real time so we can support languages without a standard writing system as well as those that are both written and spoken.

This work is built on some of the AI successes we have already had to date, such as our work to scale text-based machine translation to 101 languages by creating the first multilingual text translation system that is not English-centric.[35]

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,

[33] 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.

[34] Meta, 'Inside the Lab: Building for the Metaverse with AI', *Meta Newsroom*, 23 February 2022, https://about.fb.com/news/2022/02/inside-the-lab-building-for-the-metaverse-with-ai/.

[35] Meta AI, 'Teaching AI to translate 100s of spoken and written languages in real time', *Meta AI blog*, 23 February 2022, https://ai.facebook.com/blog/teaching-ai-to-translate-100s-of-spoken-and-written-languages-in-real-time/.

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.

These AI models, developed by Meta AI in collaboration with academic researchers at New York University's Courant Institute of Mathematical Sciences, the Universitat Politècnica de Catalunya (UPC), and the Faculty of Mathematics and the Data Science research platform at the University of Vienna, use publicly available, de-identified time series data about the spread of the disease. They have consistently been among the most accurate models since the beginning of the pandemic.[36]

Finally, Meta has most recently released Llama 2 – the next generation of our open source large language model.[37] 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. This latest release includes model weights and starting code for pre-trained and fine-tuned Llama language models — ranging from 7B to 70B parameters. Llama 2 is free for research and commercial use.

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. There has been significant demand for Llama 1 from researchers — with more than 100,000 requests for access to the large language model — and the amazing things they have achieved by building on top of it.[38] It will be exciting to see the uses that Llama 2 is put towards.

[36] Meta AI, 'Using AI to help health experts address the COVID-19 pandemic', *Meta AI blog*, 30 June 2021, https://ai.facebook.com/blog/using-ai-to-help-health-experts-address-the-covid-19-pandemic/.

[37] Meta Newsroom, *Meta and Microsoft Introduce the Next Generation of Llama,* July 2023 https://about.fb.com/news/2023/07/llama-2/

[38] *See e.g.,* Github, *Port of Facebook's LLaMA model in C/C++,* https://github.com/ggerganov/llama.cpp

Making AI more transparent and explainable

At Meta, we believe that the people who use our products should have meaningful transparency and control around how data about them is collected and used, and that this should be explained in a way that is understandable. That's 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 these systems are explainable and interpretable; and
- Investing in research, explainability and collaboration

At the user level

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

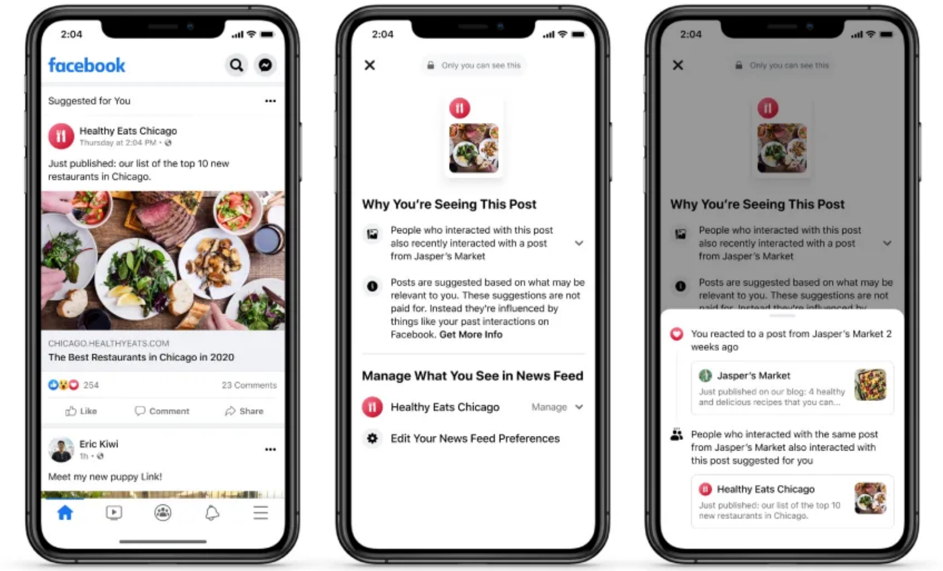
- **Why am I seeing this post?** This feature (see Figure 3 below) was launched in March 2019[39] to help people understand and more easily control what they are seeing in their Feed.

  This tool explains how a person's past interactions impact the ranking of posts in their Feed. For example, if the post is from a friend, a Group you joined, or a Page you followed, the information generally that has the largest influence over the order of posts, including: (a) how often you interact with posts from people, Pages or Groups; (b) how often you interact with a specific type of post, for example, videos, photos or links; and (c) the popularity of the posts shared by the people, Pages and Groups you follow. This was recently expanded to include context about why people are seeing suggested posts.[40]

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

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

Figure 3: Why am I seeing this post?



Healthy Eats Chicago
Thursday at 2:04 PM

Just published: our list of the top 10 new restaurants in Chicago.

CHICAGO.HEALTHYEATS.COM
The Best Restaurants in Chicago in 2020

254 · 23 Comments

Eric Kiwi
1h

Meet my new puppy Link!

Why You're Seeing This Post

People who interacted with this post also recently interacted with a post from Jasper's Market

Posts are suggested based on what may be relevant to you. These suggestions are not paid for. Instead they're influenced by things like your past interactions on Facebook. **Get More Info**

Manage What You See in News Feed

Healthy Eats Chicago

Edit Your News Feed Preferences

You reacted to a post from Jasper's Market 2 weeks ago

Jasper's Market
Just published on our blog: 4 healthy and delicious recipes that you can...

People who interacted with the same post from Jasper's Market also interacted with this post suggested for you

Healthy Eats Chicago
Just published: our list of the top 10 new restaurants in Chicago.

- **Why am I seeing this ad?** This feature allows people to see how factors like basic demographic details, interests and website visits contribute to the ads in their Feed. There are also additional details about when information on an advertiser's list matches a person's profile.[41]

  Users are able to change their Ad Preferences through the tool, if they decide that they want to take steps to ensure they don't see similar ads in future.

- **Facebook Feed Controls.** There are a number of controls that people can use to manage the content they see on Facebook, such as:
  - *Favourites:* This tool allows Facebook users to control and prioritise posts from the friends and Pages they care about most. By selecting up to 30 friends and Pages to include in Favourites, or up to 50 Instagram accounts, their posts will appear higher in ranked Facebook and Instagram Feeds respectively and can also be viewed in a separate feed populated exclusively with posts from a person's "Favourites".[42]
  - *Feed Filter Bar:* This feature allows Facebook users to alternate between different Feed experiences - the algorithmically-ranked Feed, the

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

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

    chronological Most Recent Feed[43], or the Favorites Feed discussed above.[44] In March, we announced that similar functionality is available on Instagram for both "Favourites" and "Following" (accounts that a user follows on Instagram) and both are available chronologically rather than via algorithmic ranking.[45]
  - *Facebook Feed Preferences:* This is a suite of tools that allow people to manage what they see in their Facebook Feed, including the ability to unfollow people, snooze a particular account, or prioritise Favourites.[46]

- **Instagram Feed controls.** On Instagram, the controls available to people to manage the content they see include:
  - *Recommended content controls:* this allows people to manage the types of content they want to avoid in places like Explore, Search and Reels. They can choose to hide multiple pieces of content in Explore that they are not interested in at one time. A person can also select "Not interested" on a post seen in Explore, and we will aim to avoid showing this kind of content going forward in other places where we make recommendations, like Reels, Search and more. Additionally, if a person has used the "Hidden Words" tool to avoid seeing comments or direct messages with certain words, this feature has been expanded to apply to recommended posts a person might see across Instagram. A person need only add a word or list of words, emojis or hashtags that they want to avoid – like "fitness" or "recipes" – and we work to no longer recommend content with those words in the caption or the hashtag.[47]
  - *Favourites and Following*: these tools are ways to catch up on recent posts from the accounts a person follows. The "Favourites" tool shows you the latest from accounts that a person chooses to follow, like your best friends and favourite creators. In addition to this view, posts from accounts in Favourites will also show up higher in your home feed. The "Following" tool shows posts from the accounts a person follows. Both "Favourites" and "Following" will show posts in chronological order and are available via a tap on Instagram in the top left corner of the home page.

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

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

[45] A Mosseri, 'Control your Instagram Feed with Favorites and Following', *Instagram blog*, 23 March 2022, https://about.instagram.com/blog/announcements/favorites-and-following.

[46] Facebook, *How can I see and adjust my Facebook News Feed preferences?,* https://www.facebook.com/help/371675846332829

[47] https://about.instagram.com/blog/announcements/new-ways-to-control-what-you-see-on-instagram/

At the system level

- **System Cards.** 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.

  To that end, in February 2021 we launched a prototype AI System Card tool[48] that is designed to provide insight into an AI system's underlying architecture and help better explain how the AI operates.[49]

  The System Card outlines the AI models that comprise an AI system and can help enable a better understanding of how these systems operate based on an individual's history, preferences, settings, and more. The pilot System Card we've developed, and continue to test, is for Instagram Feed ranking, which is the process of taking as-yet-unseen posts from accounts that a person follows and then ranking them based on how likely that person is to be interested in them. It is intended to be read by both experts and non-experts.

  We also published the technical paper which explains why we believe a Systems Cards approach is the best form of documentation to provide transparency of the ML systems we use and the data and models they rely on.[50]

  This past June, we expanded this prototype and shared 22 system cards that contain information and actionable insights everyone can use to understand and customise their specific AI-powered experiences in our products.[51] We released these cards to help people better understand AI's role in many Instagram and Facebook features, and to explain how people's choices and behaviours influence what content they see through our ranking and recommendation systems, such as a new video or a creator they may want to follow.

[48] See the inaugural Systems Card prototype available here: https://ai.facebook.com/tools/system-cards/instagram-feed-ranking/

[49] Meta AI, 'System Cards, a new resource for understanding how AI systems work', *Meta AI blog*, 23 February 2021, https://ai.facebook.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/.

[50] Meta AI, 'System-level transparency of machine learning', *Meta AI blog*, 22 February 2022, https://ai.facebook.com/research/publications/system-level-transparency-of-machine-learning.

[51] Meta AI, *"Introducing 22 system cards that explain how AI powers experiences on Facebook and Instagram',* June 2023 https://ai.meta.com/blog/how-ai-powers-experiences-facebook-instagram-system-cards/; system cards are now available in 22 languages in our Transparency Center, *see:* https://transparency.fb.com/features/explaining-ranking

- **Communicating the values that underpin Feed.** We provide transparency about the principles and policies that underlie the ranking algorithms used on Facebook and Instagram.

  On Facebook, we often make improvements to Feed, and when we do, we rely on a set of core values.[52] These values guide our thinking, and help us keep the central experience of Feed intact as it evolves. In summary, the values are:

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

  We published these values to provide transparency in the objectives and trade-offs that we consider in the product design behind Feed. We also continually evaluate the effectiveness of Feed ranking signals. We share updates about the biggest changes and tests we have launched on our Inside Feed blog to give people who use Facebook more control over their Feed.[53]

  Beyond sharing information about specific ranking changes, we are also making an effort to provide people with more detail about our ranking processes in general. For example, last year, the CEO of Instagram published a blog post detailing the ranking process on Instagram from start to finish.[54]

- **Policies for ranking – the Content Distribution Guidelines.** In 2021, we published the Content Distribution Guidelines to share more detail on the types of content that we demote in Facebook Feed.[55] While the Community Standards make it clear what content is removed from our services because we don't allow it, the Content Distribution Guidelines make it clear what content receives reduced distribution on Feed because it's problematic or low quality.

  The changes we make, particularly ones focused on limiting the spread of problematic content, are based on extensive feedback from our global community and external experts. Over the last few years, we've consulted more than 100

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

[53] Meta, *Inside Feed*, https://about.fb.com/news/category/inside-feed/

[54] A Mosseri, 'Shedding more light on how Instagram works', *Instagram Blog*, 8 June 2021, https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works

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

stakeholders across a range of relevant focus areas to solicit feedback on how to bring more insightful transparency to our efforts to reduce problematic content.

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

- **Responding to people's direct feedback.** We listen to people's feedback about what they like and don't like seeing on Facebook and make changes to Feed in response.
- **Incentivising creators to invest in high-quality and accurate content.** We want people to have interesting new material to engage with in the long term, so we're working to set incentives that encourage the creation of these types of content.
- **Fostering a safer community.** Some content may be problematic for our community, regardless of the intent. We'll make this content more difficult for people to encounter.

- **Policies for recommendations - the Recommendation Guidelines.** 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.

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

  To help people better understand our approach to recommendations, in August 2020, we published a set of Recommendation Guidelines, which outline the types of content that may not be eligible for recommendations.[56] 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[57] and Instagram.[58]

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

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

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

- **Transparency Center.** The Meta 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. Specifically, the Center includes an overview of how Artificial intelligence (AI) systems inform the ranking of content for many experiences on Meta's products, such as viewing Facebook Feed, watching reels on Instagram or browsing Facebook Marketplace.[59] 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.[60] 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.[61]

- **Technical research.** One of the most significant AI challenges is ensuring that AI can behave in a way that people can easily understand and be able to anticipate how others will respond to their actions. With the most widely used approach — reinforcement learning (RL), where the agents learn mainly from rewards collected during interactions with the environment — the agent typically develops its own unique behaviours and communication protocols. It might arbitrarily make decisions that are unintelligible both to humans and to other agents trained independently. This can make real world-AI collaboration difficult.

  Meta has developed a new, more flexible approach to teaching AI to cooperate and make their actions understandable to people: off-belief learning. Instead of using human labelled data, off-belief learning starts with the quest to search for a "grounded communication," where the goal is to find the most efficient way to communicate without assuming any prior conventions. To help the field of AI, we recently published a paper on our work, open-sourced the code and released a public demo where everyone can play with our model trained using off-belief learning.[62]

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

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

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

[62] Meta AI, ‘Teaching AI to be more collaborative with humans without learning directly from them’, *Meta AI blog*, 18 April 2022, https://ai.facebook.com/blog/teaching-ai-to-be-more-collaborative-with-humans-without-learning-directly-from-them/.

We have also worked with start-ups to "lift all boats" and encourage sharing best practice about AI explainability across the industry. In April 2022, we worked with cross-industry partner Trust, Transparency and Control (TTC) Labs in a series of co-creation workshops with start-ups and the Singaporean data privacy regulator to develop a framework for AI explainability. We published this framework to share our collective thinking and help to advance the debate about effective frameworks for explaining AI.[63]

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 around APAC, with recipients from Monash University and Macquarie University.[64]

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

Responsible innovation initiatives

We recognise, when working on innovative technologies, it is important to provide confidence that we are building AI in a way that is privacy-preserving and cognisant of how technology can be misused. To do this, we have published responsible innovation principles, partnered with industry and government to develop additional guidance, developed a Fairness Flow tool, and adopted an open source approach and support resources for our large language models. More detail on these efforts include:

- **Publishing principles for responsible innovation**. We released a specific five pillars of responsible AI that we have developed based on principles from the European Union and the OECD.[65] These principles are:
    - Privacy and security
    - Fairness and inclusion
    - Robustness and safety
    - Transparency and control
    - Accountability and governance.

[63] TTC Labs, *People-centric approaches to algorithmic accountability*, https://www.ttclabs.net/report/people-centric-approaches-to-algorithmic-explainability.

[64] 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/.

[65] Meta AI, 'Facebook's five pillars of responsible AIA', *Meta AI blog*, 22 June 2021, https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/.

- **Partnering with industry and government organisations.** In 2021, we launched a partnership with Business at OECD to develop a range of responsible AI case studies, including one from Meta (covering our Feed Transparency and Control efforts, and tools such as 'Why Am I Seeing This?'). The project aims to evaluate how concrete practices from business can illustrate the implementation of the OECD AI Principles, as well as highlighting some of the challenges faced in doing so. We hope it will help to facilitate a more pragmatic conversation about how to regulate such a nascent technology as AI.

  The project will complement the work of the OECD ONE.AI Expert Group on Trustworthy AI (ONE.TAI), which addresses implementation of trustworthy AI, as outlined in the OECD AI Principles.

  We are also a founding member of the Partnership on AI, which is a non-profit partnership of academic, civil society, industry and media organisations who are committed to creating solutions so AI advances positive outcomes for people and society. By convening diverse, international stakeholders, the Partnership on AI seeks to pool collective wisdom to make change. The Partnership develops tools, recommendations, and other resources by inviting voices from across the AI community and beyond to share insights that can be synthesised into actionable guidance. The Partnership then works to drive adoption in practice, inform public policy, and advance public understanding. Through dialogue, research, and education, partnerships such as these are addressing the most important and difficult questions concerning the future of AI.

- **Building in fairness by design.** One of the key teams in Meta's Responsible AI organisation is the Fairness Team, who works with product teams across the company to foster informed, context-specific decisions about how to measure and define fairness in AI-powered products. This team provides a number of tools to ensure AI systems are fair and inclusive.

  One important step in the process of addressing fairness concerns in products and services is surfacing measurements of potential statistical bias early and systematically. To help do that, we developed a tool called Fairness Flow. Using Fairness Flow, our teams can analyse how some common types of AI models and labels perform across different groups. It's important to look at fairness group by group because an AI system can perform poorly for some groups even when it appears to perform well for everyone on average. Fairness Flow works specifically by helping machine learning engineers detect certain forms of potential statistical bias in certain types of AI models and labels. It measures whether models or human-labelled training data perform better or worse for different groups of people.[66]

- **Open source approach & support**: as part of our commitment to building AI responsibly, we have adopted an open source approach with respect to our large language models that promotes transparency and access. We know that while AI has brought huge advances to society, it also comes with risk. This is also why we are providing a number of resources to help those who use Llama 2 to do so responsibly too. Specifically these are:
    - *Red-Teaming Exercises:* Our fine-tuned models have been red-teamed — tested for safety — through internal and external efforts. The team worked to generate adversarial prompts to facilitate model fine-tuning. In addition, we commissioned third parties to conduct external adversarial testing across our fine-tuned models to similarly identify gaps in performance. These safety fine-tuning processes are iterative; we will continue to invest in safety through fine-tuning and benchmarking and plan to release updated fine-tuned models based on these efforts.
    - *Transparency Schematic:* We explain our fine-tuning and evaluation methods for the model and identify its shortcomings. Our transparency schematic, which is located within the research paper, discloses known challenges and issues we've experienced and provides insight into mitigations taken and future ones we intend to explore.
    - *Responsible Use Guide:* We created this guide as a resource to support developers with best practices for responsible development and safety evaluations. It outlines best practices reflective of current, state-of-the-art research on responsible Generative AI discussed across the industry and the AI research community.
    - *Acceptable Use Policy:* We put a policy in place that prohibits certain use cases to help ensure that these models are being used fairly and responsibly.
    - *Feedback:* Meta has also created new initiatives to harness the insight and creativity of individuals, researchers, and developers around the world to get feedback on how the models

[66] Meta AI, 'How we're using Fairness Flow', to help build AI that works better for everyone', *Meta AI blog*, 31 March 2021, https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/.

Discussion of key policy issues

There are a number of policy issues identified in the Department's discussion paper, the National Science and Technology Council's Rapid Research Report on Generative AI, and in other recent reports on AI such as the Australian Human Rights Commission's report on Human Rights and Technology.

We provide some initial comments below on the key policy issues that the Department may be considering in the context of AI. But before doing so, we encourage the Department to consider what existing regulatory frameworks govern many of the policy concerns raised in the context of AI to assess the extent to which they are fit-for-purpose already or may need to be adjusted to support innovative new uses. Acknowledgement of the extent to which AI supports fulfilment of many existing regulatory requirements will also assist in a comprehensive review.

Review of existing regulatory frameworks

Before considering what additional regulatory reform may be needed to address policy concerns with respect to AI, it is helpful to review the suite of digitally focused regulatory reforms that have already been undertaken in Australia, and also the extent to which AI is supporting compliance with these requirements and supporting public policy objectives of the Australian Government and Australian community values.

The Australian Government has been active in introducing new regulation specifically related to digital platforms. In the last three years, at least 14 new digital platforms regulations have been pursued at the federal level, including online safety, privacy, consumer protection and misinformation reforms – and these are in addition to existing laws that apply such as disinformation and defamation laws.

For example, the Online Safety Act was updated recently in 2021 with underpinning regulations such as the Basic Online Safety Expectations and the development of eight mandatory industry codes as well as the Restricted Access System Declaration and the Age Verification Roadmap. These combine to contain strict requirements relating to harmful content that apply equally to content no matter whether it is generated via AI or not.

With respect to privacy, the Attorney-General's Department is undertaking a cross-economy privacy review, contemplating 116 proposals, including several related to personalisation and personalised advertising (driven by AI).

And finally, an industry code on disinformation and misinformation (instigated at the Government's request and delivered with oversight from the Australian Communications and Media Authority) was first released in 2017. Signatories have now published three annual transparency reports as part of code compliance and forthcoming draft legislation is currently being consulted on.

Much of the requirements outlined in these pieces of legislation are best fulfilled by the use of AI. For example, as outlined above, AI is integral to our integrity systems that proactively detect and action harmful content and use ranking algorithms to downrank non-policy violating but problematic content. In addition, many of these laws will apply to AI-generated content and behaviour just as they do to human-generated harmful content and behaviour.

Consideration of these existing frameworks will assist in identifying how AI regulation can be built upon existing legislation that already impacts AI , without creating tension with existing obligations.

Framework to review and identify policy issues

As part of identifying what additional regulatory frameworks or adjustments to existing regulatory models may be helpful, it is helpful to group areas of policy concern for Generative AI research models in to the following framework:

I. **Research model training data**
   A. As Generative AI is largely dependent on using web-crawled datasets, there are three issues to think about in terms of the makeup of those datasets:
      1. *Privacy:* inclusion of private/sensitive content in the dataset
      2. *Provenance:* inclusion of content from unknown sources in the dataset
      3. *Diversity:* existence of varying degrees of representation for different demographic groups within the dataset

II. **Evaluation of user inputs**
   A. As users are able to input their own queries and requests, there are a couple of things here to think about:
      1. *Integrity:* Users may ask the model to perform tasks that violate community standards or that are sensitive
      2. *Bias/stereotyping:* Users may ask/prompt the model to reproduce stereotypes

III. **Model outputs**

A. As the Generative AI is trained to create new content, there are some issues to think about with the output it produces:

1. *Misinformation:* model may inadvertently hallucinate false content
2. *Toxicity:* model may generate content that users find offensive, insensitive, or toxic
3. *Bias/stereotyping:* model may exhibit bias or stereotypes in its output without user prompting

A review of these policy concerns further identifies that many are already addressed by existing laws and/or reveals that some existing laws may need to be reconsidered to appropriately address new concerns.

For example, Generative AI can also be used to create synthetic training data, reducing the need for real data and therefore improving privacy. Synthetic data may increase the diversity of model training data and improve inclusivity in AI model performance. Generative AI may, for example, create speech data for different voices, accents, or languages that might be fit for model training. Generative AI may also be able to create synthetic data for rare languages to enable key translation benefits.

Protecting privacy in the AI context may look different than it has in the past. For example, outside of the AI context, privacy concerns might warrant not collecting sensitive information like age or gender. However, we know that including such data in training datasets is critical to ensuring that AI does not reflect biases. A lack of diverse data can lead to AI-powered outcomes that reflect problematic stereotypes or fail to work equally well for everyone.

That's why, starting in 2021, we began releasing the "Casual Conversation" dataset.[67] It's a dataset of over 45,000 videos of 3,011 participants who opted-in to explicitly provide age and gender labels themselves. It is intended to be used for assessing the performance of already trained models in computer vision and audio applications for the purposes permitted.

Principles for AI regulation

The fundamental challenge is to develop regulations that are broad and flexible enough to adapt to future technologies while not overly restrictive to the point of suppressing

[67] Meta Research, *Casual Conversations v2: Designing a large consent-driven dataset to measure algorithmic bias and robustness,* November 2022, https://research.facebook.com/publications/casual-conversations-v2-designing-a-large-consent-driven-dataset-to-measure-algorithmic-bias-and-robustness/

valuable and beneficial innovations in, and uses of, AI technology. For this reason, we need collaborative policymaking to ensure an appropriately balanced approach. Meta stands ready to collaborate with Australian policymakers on these important issues.

Many AI systems at issue are profoundly complex. There are unanswered questions about how to simultaneously achieve different policy objectives such as meaningful transparency, upholding privacy, protecting trade secrets, and encouraging innovation.

Rushing to impose onerous data, technical, and transparency legal requirements in the absence of consensus based standards and guidelines risks creating substantial risks for companies and their users.

We note that much of the Discussion Paper acknowledges this complexity and nuance. Against this background, we encourage the Department as part of any consideration of possible regulatory responses to AI to emphasise the following principles:

- **Use definitions that strike the right balance between precision and flexibility:** Any legislation should include a definition of AI that is sufficiently flexible to accommodate technical progress, but also precise enough to provide the necessary legal certainty. At the same time, however, a definition should not be too narrowly focused on a detailed and prescriptive description of the underlying technical elements of AI and machine learning because, as this is a dynamic and continuously evolving field, they will soon become outdated. We believe that legal certainty around AI developers' obligations can be achieved while still preserving the flexibility to accommodate changing needs and norms – and the ability to take full advantage of the powerful economic benefits of AI – as the technology evolves.

  We recommend adopting a definition which focuses on AI systems that learn and adapt over time because these are the capabilities that are at the core of AI, that make it different from other software applications, and that raise new and unique governance questions. Specifically, we recommend adopting a definition consistent with the definition proposed by the OECD Expert Group on AI:

  > *An AI system is a machine-based system that is capable of influencing the Environment by making recommendations, predictions or decisions for a given set of objectives. It does so by using machine and/or human-based inputs/data to: i) perceive real and/or virtual environments; ii) abstract such*

perceptions into models manually or automatically; and iii) use model interpretations to formulate options for outcomes.[68]

- **Review existing regulatory frameworks:** As the Discussion Paper acknowledges, many of the policy concerns raised in the context of AI are already addressed by existing regulatory frameworks. However, one area that it may be helpful to further consider is the extent to which AI, especially Classic AI, is helpful in meeting existing regulatory obligations for digital platforms and/or policy issues that arise via AI. Additionally, a more detailed review of existing regulatory frameworks may be helpful in assessing the extent to which they are fit-for-purpose already or may need to be adjusted, for example, as discussed in more detail above, protecting privacy in the AI context may look different than it has in the past.

- **Build upon existing legislation and adopt a suitable framework to identify policy concerns:** In an effort to be helpful to the Department as it considers what existing frameworks will address policy concerns posed by AI, and Generative AI in particular, as outlined above, we suggest that policy concerns be broken out and considered with respect to research model training data, user inputs and model outputs. This, combined with a review of existing obligations noted above, should assist in minimising tension with existing obligations, in the event that new governance frameworks are identified as being necessary. This will also help to provide greater legal clarity, avoid duplicative regulation, and ensure a proportionate approach to any novel issues.

- **Be principle-based:** Rather than codifying inflexible rules, regulators should focus on supporting and building on ongoing efforts to establish best practices in the fields of Responsible AI. Rather than prescriptive technical requirements, AI legislation should provide opportunities for stakeholders to come together to develop and regularly update the standards and best practices for assessing, measuring, and comparing AI systems as they evolve. We note that the Discussion Paper references several principles based approaches that have been adopted to date by the OECD, which provides a strong foundation on which to consider governance models.

- **Take a risk based approach that is both pragmatic and evidence-based:** The development of AI standards and regulations should be underpinned by a risk based approach, focused on the most sensitive types of AI applications and

[68] OECD Expert Group on AI, https://oecd.ai/en/ai-principles

sectors, such as in cases where AI may produce decisions that cause legal or similarly significant effects.

We note that the Discussion Paper favourably considers a risk-based approach. AI is a fast-evolving field, with new techniques emerging all the time. A risk-based approach is more future-proof than approaches that focus on particular technologies or techniques, which may become obsolete within years. In contrast, the outcomes that risk-based approaches generally seek – prevent and minimise harm, ensure protections, foster innovation – are less likely to change dramatically, even as new technologies emerge.

However, in adopting this approach it is important to carefully calibrate how to identify risk. We encourage the Department to adopt a risk-based approach that considers the technology in context, and introduces rules in a way that is proportionate to the level of risk a situation presents.[69] This reduces the likelihood that rules are introduced unnecessarily, creating barriers to innovation and adoption of useful, low-risk AI. This type of approach tends to focus on the outcomes that one wants to achieve or prevent – the 'what' – rather than how they are arrived at. This allows companies to develop their own practices, tools, and techniques to meet expectations, in comparison to more prescriptive approaches which can impose rigid processes on business models that are not well-suited to them.

Within that risk based approach, we believe that - except for exceptional, high-risk circumstances - risk assessments should be conducted by the entities (whether private or public) acting as providers for the AI system, and cover both the potential risks as well as the potential benefits of the AI systems being built and deployed. Potential legal requirements on explainability, auditing, transparency disclosures, and data subjects' right to appeal, redress, and object should only be applicable to AI applications that pose a high risk.

- **Encourage open innovation and competition:** AI should benefit everyone–not just a handful of companies. AI innovation is inevitable and it should be built by an AI research community to benefit the whole-of-society. A specific example to help illustrate this open innovation approach is Large Language Models (LLMs). LLMs. are extremely expensive to develop and train. Fostering a flourishing AI research community that enables experts from diverse disciplines to explore, challenge and

[69] See: The case of the EU AI Act: Why we need to return to a risk-based approach, IAPP

innovate with cutting-edge technology depends on democratising access to the most sophisticated models, which are mostly developed by industry.

An open innovation approach increases market contestability by spurring new market competition, creating more innovation and consumer choices. Open innovation can also facilitate new entry by providing a wide range of stakeholders with access to AI models that will allow them to innovate and compete. Open innovation also promotes sustainable economic growth by helping to close any gap by enabling researchers and SMEs to build on open source models, making new discoveries and building profitable businesses.

In addition, an open approach has safety benefits. With thousands of open source contributors working to make AI systems better, we can more quickly find and mitigate potential risks in systems and improve the tuning to prevent erroneous outputs. The more AI-related risks that are identified by a broad range of stakeholders - researchers, academics, policymakers, developers, other companies - the more solutions the AI community, including tech companies, will be able to find for implementing guardrails to make the technology safer.

The more access given to AI models, the more likely it is that toxicity and bias can be identified and appropriately addressed and mitigated.

- **Be a product of collaboration amongst multiple stakeholders:** Regulators should coordinate and collaborate with the many experts and stakeholders of the AI ecosystem and devise their legislative strategies in conjunction with other co- or self-regulatory instruments (international AI principles, standards, ethical codes of conduct, NIST AI Risk Management Framework etc). Governments, through their regulatory agencies, should explore the implementation of Regulatory Sandboxes (RS) to foster the development of new products and services within existing regulatory frameworks, and Policy Prototyping Programs (PPPs) as methods to test future laws and regulatory frameworks instruments on AI and other emerging tech. Given the difficulty in assessing the most appropriate, feasible and balanced legislative instruments on a complex topic such as algorithmic accountability, PPPs can provide a safe testing ground to assess different iterations of legislative models of governance prior to their actual enactment. An example of this type of collaboration is Open Loop,[70] a global strategic initiative by Meta that promotes and deploys experimental regulatory efforts in the field of new and emerging

[70] *Please see* https://ai.facebook.com/blog/introducing-open-loop-a-global-program-bridging-tech-and-policy-innovation/

technologies. It supports the co-creation and testing of new governance frameworks through policy prototyping programs and enables the evaluation of existing legal frameworks through regulatory sandbox exercises.

Additionally, cross-ecosystem collaboration is helpful in developing rules and norms that are globally harmonised . Globally-harmonised frameworks are necessary to ensure consistent standards around the world. Such frameworks will protect people's information wherever it goes and provide predictable rules for businesses - both being essential requirements for the long-term success of the global digital economy. Additionally, it will foster a level playing field for all AI providers operating across borders.

We note that coming out of the G7 Hiroshima Leaders Meeting in May this year, the OECD was encouraged to look at the policy impacts of generative AI and the Global Partnership of AI was tasked to conduct practical projects. These multi-stakeholder processes are good avenues for an inclusive and global discussion on the key issues pertaining to AI.

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