

Question 1

In evidence to the Committee at your appearance on 20 January 2022, Ms Garlick stated: *“Obviously we have an active online safety reform process underway at the moment with the new law taking effect on Sunday, and so we have teams set up to identify our compliance approach to that, and similarly with privacy law reform we have a team set up. And so where there are very specific laws that are designed to ensure that digital platforms are meeting Australian standards absolutely we have compliance teams that are looking at that...”*

- a. Does Facebook have a dedicated team to proactively ensure compliance with the Commonwealth Racial Discrimination Act (particularly Section 18C)?
- b. Does Facebook have a dedicated team to proactively ensure compliance with state based vilification laws?

Meta response:

We have dedicated teams that work to develop and update our policies with respect to hateful content that targets people based on race.¹ We also have operations teams that work to review user reports and content identified by our classifiers to further train our proactive detection technology. We have shared details about our work to combat hate speech on our services, including our significant investment in tools to proactively detect hate speech, in our submission.²

Much of our work to prevent the sharing of hateful content that targets people based on protected characteristics - including race - means a significant amount of content prohibited by laws such as the Commonwealth Racial Discrimination Act and state vilification laws is already prohibited by our global policies. When we receive a request to restrict access to content from the Australian Human Rights Commission or other key stakeholders, we first review it against our global policies. If the content violates those policies, we remove it from our services.

If the content does not violate those policies, legal review will be carried out. If the request is legally valid, and the content is locally unlawful, we generally restrict access to

¹ See Meta, *Community Standards - Hate Speech*

<https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/>

² Meta, ‘Submission to the Select Committee on Social Media and Online Safety - submission 49’, *Parliament of Australia*,

https://www.aph.gov.au/Parliamentary_Business/Committees/House/Social_Media_and_Online_Safety/SocialMediaandSafety/Submissions

the content in the country where it is alleged to be unlawful. We are transparent about the content we restrict based on local law in our Transparency Centre.³

³ See Meta, *Content Restrictions Report*: <https://transparency.fb.com/data/content-restrictions/>

Question 2

There were a flurry of media reports last year that suggested Facebook was considering shutting Crowdtangle down. Were these reports correct?

Meta response:

We have recently announced a new data and transparency team, inclusive of CrowdTangle. For those who use CrowdTangle, there has been no change in the day-to-day functioning of the product experience. We will continue to evaluate how to best promote transparency on our products and services.

Question 3

3. Mr. Brandon Silverman, the founder of the crowdtangle before it was acquired by Facebook recently left the company, telling the NY Times

“(Facebook) gave us a lot of freedom and resources and support to do this work for four years when a lot of platforms were doing nothing.. There was a vision about transparency that I believed in and my team had come to believe in that (but) it was clear we wouldn’t be able to pursue inside Facebook as much as we had in the past”

Has Facebook’s attitudes towards transparency changed in the way that Mr Silverman describes?

Meta response:

We can’t speak to Mr. Silverman’s opinion or his personal recollections of his time at Meta, but Meta has objectively become *more* transparent over the years, not less.

Meta has steadily increased the transparency around our policies, enforcement and processes since 2014. For example, we started sharing information about content restrictions and responses to law enforcement requests in 2014; began sharing data about our Community Standards enforcement work in March 2018, which are now released quarterly; and we started sharing our Ad Library Report in 2020 (in Australia).

In May 2021, we launched a new Transparency Center⁴ to provide a hub for all our integrity and transparency work. In addition to information on how we enforce our Community Standards, the Transparency Center has also become a central destination for all updates on how Meta is responding to decisions, recommendations and most case updates from the Oversight Board.

We welcome transparency as part of ensuring that companies such as Meta are held accountable, and we also believe that data and transparency can contribute to important public policy discussions around public policy.

We continue to look at instances where we can improve. For example, we have internally created a new data and transparency team that will be operating with a cohesive approach for our transparency efforts. We’re also developing and will continue to evaluate

⁴ Meta, *Transparency Centre*, <https://transparency.fb.com/oversight/>

a more comprehensive strategy for how we build on some of these transparency efforts in future.

Question 4

In the same NY Times article, Brian Boland, a Facebook vice president who was Mr. Silverman's boss before resigning in 2020 is quoted as saying the CrowdTangle data *"told a story (Facebook) didn't like and frankly didn't want to admit was true."* The article asserts that Mr Silverman's team was then disbanded – is this true?

Meta response:

We simply disagree with Mr Boland. The internal reorganisation of our data and transparency teams, including CrowdTangle, was done to better integrate them together for a cohesive approach. Our record to date demonstrates that we remain committed to transparency, including data transparency.

Question 5

Your submission talks about your third-party factchecking program that enables AAP and AFP to review and rate the accuracy of posts on Facebook and Instagram before publicly posting their factchecks on their website.

What evidence do you have that Factchecking works in remedying the harms caused by mis and disinformation?

Meta response:

Although the question refers to disinformation, there's a noteworthy distinction between disinformation and misinformation. Disinformation involves sharing content with the deliberate intent to mislead as part of a coordinated manipulation campaign or information operation. More information about our work in this space can be found in our Coordinated Inauthentic Behaviour reports.⁵ Our work to combat disinformation is separate from our fact-checking program, which is designed to assist with combatting misinformation.

Misinformation, as distinct to disinformation, is about the content itself: false or misleading information. We rely on our global network of fact-checkers to review and rate the accuracy of potential misinformation that doesn't violate our Community Standards. When a fact-checker rates a piece of content as false, we significantly reduce its distribution so that fewer people see it. We notify people who try to share the content – or previously shared it – that the information is false, and we apply a warning label that links to the fact-checker's article disproving the claim.

We know that our fact-checking program is making a difference– more than 95% of the time when people see one of our fact-checking labels, they don't go on to view the original content.⁶

A broad range of academic literature has confirmed the impact that fact-checking can have.

A selection of examples of research into fact-checking is provided below.

⁵ Meta, 'Recapping our 2021 coordinated inauthentic behaviour enforcements', 20 January 2022, *Meta Newsroom*, <https://about.fb.com/news/2022/01/december-2021-coordinated-inauthentic-behavior-report/>

⁶ G Rosen, 'How we're tackling misinformation across our apps, 22 March 2021, *Meta Newsroom*, <https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/>

Research from Ethan Porter (George Washington University) and Thomas Wood (Ohio State University) found that across 52 studies, fact-checking corrections reduce the belief in misinformation across the ideological spectrum.⁷

Additional experiments conducted in Argentina, Nigeria, South Africa, and the United Kingdom found that fact-checking does result in an enduring, statistically-significant increase in accurate beliefs.⁸ The researchers concluded that this “evidence underscores that fact-checking can serve as a pivotal tool in the fight against misinformation.”

Research from Brendan Nyhan (Dartmouth University) and Jason Reifler (University of Exeter) have also confirmed that exposure to fact-checking reduces misperceptions amongst users.⁹

Additionally, in 2020, dozens of the world’s top scientists studying misinformation came together and published the Debunking Handbook, consolidating the results of dozens of scientific studies on the issue.¹⁰ They conclude that “fact-checking can reduce people’s beliefs in false information” and recommend “debunk[ing] often and properly.” Notably, the authors address the criticisms that while fact-checking may change belief, it does not change behavior. They write that one should not “refrain from debunking because you are worried that it will not change behaviour. Successful debunking can affect behaviour—for example, it can reduce people’s willingness to spend money on questionable health products or their sharing of misleading content online.”

We are also specifically funding Australian research relating to fact-checking. Last year, we announced the successful application of research funding by Andrea Carson, James Meese, Justin B. Phillips, Leah Ruppanner, La Trobe University for ‘How fact checkers compare: News trust and COVID-19 information quality’.¹¹

⁷ T Wood & E Porter, ‘The elusive backfire effect: mass attitudes’ steadfast factual adherence’, 31 December 2017, SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2819073

⁸ E Porter & T Wood, ‘The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom’, 14 September 2021, *Proceedings of the National Academy of Sciences of the United States of America*, <https://www.pnas.org/content/118/37/e2104235118>

⁹ B Nyhan & J Reifler, ‘The roles of information deficits and identity threat in the prevalence of misperceptions’, 6 May 2018, *Journal of Elections, Public Opinions and Parties*, Vol. 29, 2, <https://www.tandfonline.com/doi/abs/10.1080/17457289.2018.1465061>

¹⁰ George Mason University, *The Debunking Handbook 2020*, <https://www.climatechangecommunication.org/debunking-handbook-2020/>

¹¹ Meta, ‘Announcing the 2021 recipients of research awards in misinformation and polarisation’, 14 September 2021, *Meta Research*, <https://research.facebook.com/blog/2020/08/announcing-the-winners-of-facebooks-request-for-proposals-on-misinformation-and-polarization/>

There has also been research to indicate benefits resulting from Meta's work to combat misinformation in general.

- First, Alcott, Gentzkow and Yu published a study on misinformation on Facebook and Twitter.¹² The researchers began by compiling a list of 570 sites that had been identified as false news sources in previous studies and online lists. They then measured the volume of Facebook engagements (shares, comments and reactions) and Twitter shares for all stories from these 570 sites published between January 2015 and July 2018. The researchers found that on Facebook, interactions with these false news sites declined by more than half after the 2016 election, suggesting that “efforts by Facebook following the 2016 election to limit the diffusion of misinformation may have had a meaningful impact.”
- A University of Michigan study on misinformation had similar findings about the effectiveness of our work.¹³ The Michigan team compiled a list of sites that commonly share misinformation by looking at judgements made by two external organizations, Media Bias/Fact Check and Open Sources. Because this categorisation is based on somewhat “imprecise criteria and fallible human judgments,” the researchers lightheartedly refer to these sites as “Iffy” sites and have coined a metric called the “Iffy Quotient” to measure how much content from those sites has been distributed on Facebook and Twitter.

The Iffy Quotient for Facebook spiked in 2016, leading up to the US election, but improved beginning in mid-2017. When an “engagement-weighted” version of the Iffy Quotient is considered — that is, when social media interactions like likes, comments, and shares are factored in — the study finds that Facebook now has 50% less “Iffy Quotient content” than Twitter and has returned to its early 2016 levels. The researchers cite some of our recent efforts, noting that, “Facebook may have been more successful at detecting and countering fake accounts and manipulation campaigns, more aggressive in discounting ranking signals that are associated with Iffy sites, or more aggressive in demoting particular articles and sources.”

¹² H Alcott, M Gentzkow & C Yu, ‘Trends in the diffusion of misinformation on social media’, October 2018, *Meta Research*, <https://about.fb.com/wp-content/uploads/2018/10/fake-news-trends.pdf>

¹³ P Resnick, A Ovadya & G Gilchrist, ‘Iffy quotient: a platform health metric for misinformation’, 10 October 2018, *School of Information Centre for Social Media Responsibility*, vol. 1, <https://about.fb.com/wp-content/uploads/2018/10/UMSI-CSMR-Iffy-Quotient-Whitepaper-810084.pdf>

- A new study conducted by researchers at the University of Michigan, Princeton University, University of Exeter and Washington University at St. Louis offers encouraging findings about the scale and spread of misinformation since the 2016 US elections.¹⁴ Namely:
 - *Fake news exposure fell dramatically from 2016 to 2018.* The researchers found that there was a substantial decline (75%) in the proportion of Americans who visited fake news websites during the 2018 midterm elections, relative to the 2016 elections.
 - *Also during the 2016 – 2018 period, Facebook’s role in the distribution of misinformation was dramatically reduced.* To determine Facebook’s role in spreading false news, the researchers looked at the three websites people visited in the 30 seconds before arriving at a fake news site. Between the fall of 2016 and the summer and fall of 2018, Facebook’s role in referring visits to fake news sites had dramatically dropped.

We are convening two dedicated workshop series globally with leading misinformation, safety and expression experts to advise us on best practice research and thinking on effectiveness of misinformation enforcement (including fact-checking). One of those is focused on the Asia-Pacific region, including Australia.

¹⁴ T Lyons, ‘New research shows Facebook making strides against false news, 19 October 2018, *Meta Newsroom*, <https://about.fb.com/news/2018/10/inside-feed-michigan-lemonde/>

Question 6

What is the average time that elapses between the publication of a fact check by AAP or AFP and the original posting of that content to Facebook?

Meta response:

It's not possible to give a timeframe around how long it takes a fact checker to verify content after it is posted on Facebook. This is because content is flagged to fact checkers in a variety of ways, and it is at the discretion of the fact checkers as to which pieces of content they review and rate. The amount of time it takes a fact checker to verify a claim and undertake a fact check can also vary, depending on the complexity of the claim they are reviewing.

Content is flagged to our third party fact checkers to review in several ways:

- our third-party fact-checkers proactively identify the content themselves
- our technology identifies potential false stories for third-party fact-checkers to review. For example, when people on Facebook submit feedback about a story being false or comment on an article expressing disbelief, these are signals that a story should be reviewed
- we also have a similarity detection system that helps us identify more debunked content than what our fact checkers see.

What's important is once a story is debunked by our third party fact checkers, we make it less visible on Facebook and Instagram.

Artificial intelligence plays an important role in helping scale the efforts of our third party fact checkers. After one fact check on one piece of content, we're able to kick off similarity detection which helps us identify duplicates of debunked stories, and reduce their distribution.¹⁵

These new posts are then fed back into the machine learning model which helps improve its accuracy and speed.

¹⁵ T Lyons, 'Increasing our efforts to fight false news', 21 January 2018, *Meta Newsroom*, <https://about.fb.com/news/2018/06/increasing-our-efforts-to-fight-false-news/>

Question 7

How is the subject of the content to be fact checked by AAP or AFP identified? Can individual members of the public or civil society groups request factchecks?

Meta response:

Fact checkers can identify hoaxes based on their own reporting, and Meta also surfaces potential misinformation to fact-checkers using signals, such as feedback from our community or similarity detection. During major news events or for trending topics when speed is especially important, we use keyword detection to gather related content in one place, making it easy for fact checkers to find. For example, we've used this feature to group content about COVID-19, elections, natural disasters, conflicts and other events.

Fact-checkers prioritise:

- viral false information.
- hoaxes that have no clear basis in fact.
- content presented as opinion, but based on underlying false information.
- provably false claims, especially ones that are timely, trending and important to the average person.

Our technology can detect posts that are likely to be misleading based on various signals, including how people are responding and how fast the content is spreading. We then surface these posts to fact checkers. Signals that help us identify false information include:

- Comments on posts that express disbelief.
- Machine learning models that continuously improve our ability to predict false information.

Facebook and Instagram users can also report potential misinformation using in-app reporting tools.

Anybody can send material to our third party fact-checking partners and request that they review it. Information about how to refer content to AAP and AFP can be found at:

AAP: <https://factcheck.afp.com/contact>

AAP: <https://www.aap.com.au/make-a-submission/>

It is at the discretion of the fact checkers which content they choose to review and rate.

Question 8

How many pieces of content are AAP and AFP able to fact check within a four week period? What are the constraints on their capacity to fact check content? How many fact checks of Facebook content per four month period are AAP and AFP funded to undertake?

Meta response:

We pay partners based on the number of fact checks they submit to us, but specific details of our financial agreements are commercial-in-confidence. Our partners post articles on content they fact check on their website. Every fact check article submitted can be applied to many pieces of content on our platform. In 2021, we confirmed that we've added warning labels to more than 190 million pieces of COVID-19 content globally thanks to our network of fact-checking partners.¹⁶

You can review their reports here:

- AFP Australia: <https://factcheck.afp.com/afp-australia>
- AAP: <https://www.aap.com.au/factcheck/>

Since we want our fact-checking partners to focus as much of their time as possible on original reporting, we have systems in place to find both identical and similar content:

- Identical content: When fact checkers rate a video or image, we're able to find near-exact duplicates automatically, and label them.
- Similar content: When fact checkers submit an article to us, we then run this through matching models, in order to surface more content to partners that might make the same claim. This helps them debunk a higher volume of content, more efficiently.¹⁷

Our work with third-party fact checkers is not just meant to educate people about what has been disputed; it also helps us better understand what might be false and show it lower in News Feed. False ratings from third-party fact-checkers are a helpful signal that we use to inform our machine learning models, so that we can more quickly and accurately detect future false stories. This means that over time we're getting smarter and faster in determining what articles might be hoaxes and sending them to fact checkers to review.

¹⁶ G Rosen, 'Community standards enforcement report: second quarter 2021', 18 August 2021, *Meta Newsroom*, <https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021/>

¹⁷ T Lyons, 'Increasing our efforts to fight false news', 21 January 2018, *Meta Newsroom*, <https://about.fb.com/news/2018/06/increasing-our-efforts-to-fight-false-news/>

Question 9

Your submission says that Facebook then removes or deprioritises material that has been identified as misleading through this factchecking process. Does this treatment only apply to identical content?

Or does it capture similar misleading claims made in different forms on Facebook? How do you measure your success in deprioritising similar misleading claims made in different forms?

Meta response:

We have systems in place to find both identical and similar content:

- **Identical content:** When fact checkers rate a video or image, we're able to find near-exact duplicates automatically, and label them.
- **Similar content:** When fact checkers rate a piece of content, we then run this through matching models, in order to surface more content to partners that might make the same claim. This helps them debunk a higher volume of content more efficiently, and slow down the spread of misinformation.¹⁸

Artificial intelligence plays an important role in helping scale the efforts of our third party fact checkers. After one fact check on one piece of content, we're able to kick off similarity detection which helps us identify duplicates of debunked stories, and reduce their distribution.¹⁹

These new posts are then fed back into the machine learning model which helps improve its accuracy and speed in detecting potential misinformation.

When fact checkers rate a piece of content, our matching models also find other content that might make the same claim, and we surface them to the fact checkers while reducing the virality of the content. This helps them debunk a higher volume of content more efficiently, and slow down the spread of misinformation.

¹⁸ T Lyons, 'Increasing our efforts to fight false news', 21 January 2018, *Meta Newsroom*, <https://about.fb.com/news/2018/06/increasing-our-efforts-to-fight-false-news/>

¹⁹ Ibid.

Question 10

Your submission states that

“When pages... repeatedly share content that’s been debunked by fact-checking partners, they will... lose the ability to advertise or monetise within a given time period.”

How much has the United Australia Party spent on advertising on Facebook?

Has the UAP ever lost the ability to advertise on Facebook?

Meta response:

Meta’s Ad Library²⁰ contains publicly available information about expenditure of advertising on Facebook or Instagram relating to political or social issue ads.

The Ad Library includes all active ads any Page is running, along with more Page information such as creation date, name changes, Page merges and the primary country of people who manage Pages with large audiences. The information is now available to everyone through the Ad Library, including people who aren’t on Facebook.

We also have reporting functionality so that anybody can compare the amount spent on advertising by particular Pages.

Current figures for all political parties, including for the United Australia Party, can be found in the Ad Library.

For details about the enforcement history behind any particular Page, we recommend you contact the administrators of that Page.

²⁰ Meta, *Meta Ad Library*,
https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=US&media_type=all

Question 11

In November, you announced that you would delay implementing end to end encryption on messages on Facebook and Instagram from 2022 until 2023.

Why did you announce this decision to implement end to end encryption on your messaging services without a proper plan to prevent child abuse going undetected on its platforms in the first place?

Meta response:

Claims that Meta does not have plans to keep our community safe when Messenger moves to encryption are untrue.

End-to-end encryption helps to protect the safety and security of private messaging users. Getting safety right has always been part of our plan, as illustrated in our original announcement.²¹ We've always said this would be a technically complicated, long-term project, and we're taking our time to get this right.

In December last year, we published a blog post²² which outlines our three part strategy:

- Working to prevent abuse from happening in the first place,
- Giving people more controls to help them stay safe and
- Responding to reports on potential harm.

Preventing abuse

Preventing abuse from happening in the first place is the best way to keep people safe. In an end-to-end encrypted environment, we can use artificial intelligence to proactively detect accounts engaged in certain malicious patterns of behavior even without scanning people's private messages. Our technology will look across non-encrypted parts of our platforms — like account information and photos uploaded to public spaces — to detect suspicious activity and abuse.

For example, if an adult repeatedly sets up new profiles and tries to connect with minors they don't know or messages a large number of strangers, we can intervene to take action, such as preventing them from interacting with minors. We can also default minors

²¹ M Zuckerberg, A privacy focussed vision for social networking, 12 March 2021, *Facebook*, <https://www.facebook.com/notes/2420600258234172/>

²² A Davis, 'Our approach to safer private messaging', 1 December 2021, *Meta Newsroom*, <https://about.fb.com/news/2021/12/metas-approach-to-safer-private-messaging/>

into private or “friends only” accounts. We’ve started to do this on Instagram and Facebook.

We also educate young people with in-app advice on avoiding unwanted interactions. We’ve seen tremendous success with our safety notices on Messenger, which are banners that provide tips on spotting suspicious activity and taking action to block, report or ignore/restrict someone when something doesn’t seem right.²³ We developed these safety tips using machine learning to help people avoid scams, spot impersonations and, most urgently, flag suspicious adults attempting to connect to minors. And, this feature works with end-to-end encryption.

Responding to potential harm

In addition to the work that we do to proactively detect abuse on our services, reporting is an essential tool for people to stay safe and help us respond to abuse effectively. We’re making it much easier to report harm and educating people via Safety Notices in Messenger. We also recently made it easier to report content for violating our child exploitation policies.²⁴ People can select “involves a child” as an option when reporting, which helps us address violating content quicker. Our goal is to encourage significantly more reporting by making it more accessible, especially among young people. As a result, we’re seeing close to 50% year-over-year growth in reporting, and we’re taking action to keep Messenger and Instagram DMs safe.

We’ll continue to enforce our Community Standards on Messenger and Instagram DMs with end-to-end encryption. Reporting allows us to see portions of the conversation that were previously unavailable to us so that we can take action if violations are detected — whether it’s scams, bullying, harassment or violent crimes. In child exploitation cases, we’ll continue to report these accounts to NCMEC.²⁵ Whether the violation is found on or through non-encrypted parts of our platform, or through user reports, we’re able to share data like account information, account activity and inbox content from user reported messages.

We also want to educate more people to act if they see something and avoid sharing harmful content, even in outrage. We have begun sending alerts informing people about the harm that sharing child exploitation content, even in outrage, can cause by warning

²³ J Sullivan, Preventing unwanted contacts and scams in messenger, *Messenger News*, 21 May 2020, <https://messengernews.fb.com/2020/05/21/preventing-unwanted-contacts-and-scams-in-messenger/>

²⁴ A Davis, ‘Preventing child exploitation on our apps’, 23 February 2021, *Meta Newsroom*, <https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/>

²⁵ Meta, ‘How Meta works with law enforcement’, 19 January 2022, *Transparency Centre*, <https://transparency.fb.com/en-gb/policies/improving/working-with-law-enforcement/>

them that it's against our policies and will have legal consequences. We'll continue to share these alerts in an end-to-end encrypted environment, in addition to reporting this content to NCMEC. We've also launched a global "Report it, Don't Share it" campaign reminding people of the harm caused by sharing this content and the importance of reporting this content. Our research,²⁶ based on a NCMEC supported taxonomy, leads us to estimate that more than 75% of people sharing CSAM did not exhibit malicious intent (i.e. did not intend to harm a child). Instead, they appeared to share for other reasons, such as outrage or in poor humor (i.e. a child's genitals being bitten by an animal).

Even in the context of encrypted systems, there is data we can provide to law enforcement to investigate when requested via valid legal process, such as who users contact, where they were when they sent a message and when they sent it.

For example, WhatsApp relies on all available unencrypted information, including user reports, to detect and prevent this kind of abuse, and we are constantly improving our detection technology.

WhatsApp's detection methods include the use of advanced automated technology, including photo- and video-matching technology, to proactively scan unencrypted information such as profile and group photos and user reports for known CEI. We have additional technology to detect new, unknown CEI within this unencrypted information. We also use machine learning classifiers to both scan text surfaces, such as user profiles and group descriptions, and evaluate group information and behavior for suspected CEI sharing.

Along with our proactive detection work, WhatsApp encourages users to report problematic content to us. Users can also block or report an individual account or group at any time.

Using these techniques, WhatsApp made 400,000 reports to NCMEC in 2020 without breaking encryption.²⁷

²⁶ A Davis, 'Preventing child exploitation on our apps', 23 February 2021, *Meta Newsroom*, <https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/>

²⁷ See W Cathcart, *Twitter*, 7 August 2021, <https://twitter.com/wcathcart/status/1423701475595755524>

Question 12

The Department of Home Affairs submission has suggested that

“The Department’s engagements with Meta and other companies with ‘privacy first’ policies reveal a degree of seeming indifference to public safety imperatives, including in relation to children. For example, end-to-end encryption provides limited advantages over and above network level encryption. In the case of Facebook Messenger for example, end-to-end encryption will only apply to the content of messages, which has less commercial value to the company. The Department understands that personal data, such as metadata and site and cookie tracking, could still be exploited by Meta for commercial purposes, in line with their business model.”

Do you want to respond to this assertion?

Meta response:

The assertion that we are indifferent to public safety imperatives is not correct.

Our track record of industry leading safety measures, as outlined in our submission²⁸ to the Committee, demonstrates our firm commitment to investing in safety and security. Our answer to Question 11 provides more information about the steps we are taking to protect the safety of our users in private messaging.

²⁸ Meta, ‘Submission to the Select Committee on Social Media and Online Safety - submission 49’, *Parliament of Australia*, https://www.aph.gov.au/Parliamentary_Business/Committees/House/Social_Media_and_Online_Safety/SocialMediaandSafety/Submissions

Question 13

The Department of Home Affairs said

“in almost two years since tech companies endorsed the Voluntary Principles [to Counter Online Child Sexual Exploitation and Abuse], there is limited evidence as to the degree of implementation and the level of success”.

What evidence can you offer of your implementation of these voluntary principles?

Meta response:

The development and promotion of the Voluntary Principles was led - and is still being led - by Five Eyes countries with the support of WePROTECT. We and several other tech companies pledged our support for these principles, as did the Technology Coalition (of which Meta is a member). Additionally, the digital industry via the Tech Coalition have taken several measures to drive awareness of the principles and help other companies put them into practice, including developing a guide for companies on how to operationalise the principles and doing live trainings for other members of industry.

Following the publication of the Voluntary Principles, the Technology Coalition announced Project Protect, a renewed commitment and investment to protect kids online and guide its work for the next 15 years. As part of this initiative, the Technology Coalition announced five pillars of work, many of which track to the goals of the Voluntary Principles. Specifically, the Technology Coalition has stood up a transparency effort to help track industry efforts and progress in its efforts to thwart child sexual exploitation online.

For its part, Meta provides information to Australian policymakers about our efforts to combat online child sexual exploitation and abuse. The most comprehensive summary is included in our recent submission to the Parliamentary Joint Committee on Law Enforcement.²⁹

²⁹ Meta, 'Submission to the Parliamentary committee Committee on Law enforcement capabilities in relation to child exploitation - submission 24, *Parliament of Australia*, https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Law_Enforcement/ChildExploitation/Submissions

Question 14

During Dr Salter's appearance before this committee, he discussed an open letter to Mr Zuckerberg sent by 59 child protection campaigners and experts calling on Facebook to take five steps to improve the safety of its platform. Those five steps were to: share all of its internal research on the impact its platforms have on children's wellbeing; set out what research has been conducted on how the company's services contribute to child sexual abuse; publish risk assessments of how its platforms affect children; provide details of an internal reputational review of its products; and review the child protection implications of encrypted messaging.

In talking about this letter, Dr Salter told the Committee:

First and foremost, what we're asking for is just basic risk assessment practices and transparency. What are the sorts of internal processes and projects that Facebook has put in place in order to assess the risk to children for some of the major structural changes that they're proposing to make to their services? That includes end-to-end encryption for Messenger, which we're really worried about. It also includes potential changes and alterations to Instagram. What does Facebook know internally about the risk and impact to children posed by their platforms? What is the child protection evidence base behind some of their proposed initiatives?

And we're asking for transparency. Far too often, what we're provided with from social media companies in terms of their reports is simply what they choose to release to us. They define the problem. They define the terms in the way that is most suitable to them, and they release the statistics that are most friendly to them. So we're asking for simple and basic transparency and accountability. We're asking for their evidence base for the changes that they want to make to their servers. Given what is at stake and what is at risk here, which again is the sexual exploitation of children, it seems to me that this is a very low bar for them to clear, and they have made no commitment whatsoever in order to do that.

What is Facebook's response to this open letter and Dr Salter's evidence?

Meta response:

Meta provides a significant amount of transparency about the work we do on online safety. We have provided a detailed response to the open letter, where we have outlined that we share more information with researchers and academics than any other platforms.

We have already contributed to more than 300 peer reviewed articles in the last year,

but we want to be more transparent about the research we do, both internally and in collaboration with external researchers.

We're working through how we can allow external researchers more access to our data in a way that respects people's privacy. We recently announced the pilot launch of a new tool called the Researcher API, which was specifically designed for academic needs.³⁰ It equips qualified academics to conduct longitudinal research across all public Facebook Pages, Groups, Posts and Events in the US and select EU countries. Researchers can use this product to understand how public discussions on Facebook influence the social issues of the day. We offer this product via the Researcher Platform, which allows us to share privacy-protected data in a secure way.

We have invited a small group of qualified academics to test this product and provide feedback so we can iterate and improve it, before launching to a broader group of researchers.

However, sharing personal data about our users with academics can only be done in accordance with our obligations under privacy and data protection law.

³⁰ Meta, *Researcher API*, <https://fort.fb.com/researcher-apis>

Question 15

The Department of Home Affairs Submission to this inquiry stated:

“The world-leading innovation demonstrated by many digital platforms in developing their products and services has not been evident when it comes to addressing user safety. While not alone, amongst the “big tech” companies, Meta is frequently the most reluctant to work with Government to promote a safe online environment, adopt a safety-by-design approach and take adequate proactive measures to prevent online harms.”

Why do you believe the Department has singled you out in this way?

Meta response:

Meta has long taken an industry-leading approach to safety and security on our platforms. For example, we have invested more than US\$13 billion (~AU\$18 billion) on safety and security since 2016, and we spent approximately US\$5 billion (~AU\$6.9 billion) in 2021 alone.

We were the first company to publicly endorse the eSafety Commissioner’s safety by design principles.³¹ We continue to work closely with the eSafety Commissioner’s Office on later phases of their safety by design work.

We have also had extensive engagement with the Department of Home Affairs in a number of fora. Some examples include:

- We engaged closely with the Government (via the Department of Home Affairs) via the OECD process to develop a voluntary transparency reporting framework on terrorist and violent extremist content (the only technology company to co-chair one of the working groups in that work stream).
- We have proactively initiated and continued a close working relationship with the Counter Foreign Interference Coordinator and relevant teams working on foreign interference in advance of the Australian federal election.
- We are in regular and close contact with the areas of the Department on extremism, disinformation and social cohesion in relation to any content on our services they would like us to review.
- We have been in close contact with the Department in the lead-up to the conclusion of an agreement between Australia and the United States under the US

³¹ Safety by Design Youth Jam, *Facebook*, August 2019, <https://www.facebook.com/MetaAustralia/videos/910843179301219/>

CLOUD Act. Now the agreement has been signed, we look forward to continuing that close contact to implement systems in response to the agreement.

- And we have engaged extensively over many years with the Department of Home Affairs to openly share information about our thinking on safety measures in an end-to-end encrypted environment and to seek their feedback.

In addition to our work with the Department of Home Affairs, we work with a number of different parts of the Australian Government on online safety, in particular the Office of the eSafety Commissioner and law enforcement agencies. We would welcome any opportunities to work together even closer.

Any questions about the motivation of the Department of Home Affairs in making this claim should be directed to the Department.

Question 16

The Department of Home Affairs submission asserts that

“The Department’s activities to limit the spread of terrorist violent and extremist content (TVEC) online includes working with major platforms to encourage the proactive identification and removal of extremist content through operational assistance, policy development and legislative obligation, including through the Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019 and the Online Safety Act 2021. The Department leads Australia’s representation and participation on international forums and industry groups relating to TVEC on the internet, including the Global Internet Forum to Counter Terrorism. Complementary to work done by the eSafety Commissioner, the Department identifies and refers TVEC to digital platforms for consideration against their terms of service for removal.”

How frequently has the Department of Home Affairs shared TVEC content with Facebook this year?

Meta response:

Meta was an active member of the Australian Government Taskforce on Terrorist and Extreme Violent Material formed in 2019 to encourage greater collaboration between the Australian Government and the technology sector. Consistent with the final report of that Taskforce, we have submitted two annual reports on our work to combat the spread of terrorist violent and extremist content (TVEC) online to the Department of Infrastructure, Transport, Regional Development and Communications, as required by the Taskforce’s final report. We understand the Department of Home Affairs is now responsible for this process but have been advised that a third annual report is not necessary.

Under the Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019 and the Online Safety Act 2021, we work closely with the Office of the Australian eSafety Commissioner and federal and state police. However, we will respond and action (as necessary) concerns and complaints raised with us by any stakeholder, including the Department of Home Affairs.

We provided more detail about our work to combat the sharing of terrorist violent and extremist content on our services in our submission and appearance before the Joint Committee on Law Enforcement’s recent inquiry³² on the Criminal Code Amendment

³² Meta, ‘Submission to the Joint Committee on Law Enforcement, Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act - submission 14’, *Parliament of Australia*, https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Law_Enforcement/AVMAAct/Submissions

(Sharing of Abhorrent Violent Material) Act 2019. In our submission to that inquiry, we shared that:

“We’ve worked with eSafety on a small number of incidents: our efforts to work with the Australian Government have been much greater than just in relation to receiving notices. We have established a working relationship of informally briefing eSafety (at a minimum) whenever we see possible extreme violent, terrorist or crisis content on our services that may be of interest to them. We have also proactively notified the AFP of a number of instances where we have seen content on our services that could potentially constitute AVM.”³³

The Department of Home Affairs has not shared any intelligence specifically about dangerous organisations or individuals with Meta (noting that there are referrals on other issues such as social cohesion and misinformation).

³³ Meta, ‘Submission to the Joint Committee on Law Enforcement, Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act - submission 14’, Page 7, *Parliament of Australia*, https://www.aph.gov.au/Parliamentary_Business/Committees/Joint/Law_Enforcement/AVMAAct/Submissions

Question 17

What practical cooperation has the Department of Home Affairs provided to you on detecting and removing TVEC? Have they shared intelligence? Have they identified dangerous groups/individuals to you?

Meta response:

We have dedicated channels for governments and law enforcement to contact us, and we support their investigations in accordance with our terms of service and applicable law.

Meta receives reports of TVEC from a range of sources including law enforcement, researchers and community organisations. We also undertake significant proactive work ourselves.

Meta designates non-state actors under our Dangerous Individuals and Organisations policy³⁴ after a rigorous process that takes into account both online and offline behaviour. Through our relationships with law enforcement agencies, researchers, reporters and government, as well as our own investigations, we receive information about offline behaviour of dangerous groups and individuals that help inform our decision making process around designations.

The Department of Home Affairs has not shared any intelligence specifically about dangerous organisations with Meta (noting that there are referrals on other issues such as social cohesion and misinformation).

³⁴ See Meta, *Community Standards - Dangerous Individuals and Organisations*, <https://transparency.fb.com/en-gb/policies/community-standards/dangerous-individuals-organizations/>

Question 18

What more could the Department of Home Affairs be doing more in this space?

Meta response:

We welcome increased collaboration and cooperation with all relevant parts of the Australian Government. We work closely with the respective parts of the Home Affairs Department to share information about our work on areas such as:

- Terrorist and extreme violent material
- Child sexual abuse material
- Foreign interference and electoral interference
- Misinformation
- Social cohesion and counterspeech
- Electronic surveillance
- Implementation of the US-Australia CLOUD Act Agreement
- What tech companies' work with law enforcement can look like in relation to end-to-end encrypted services.

Our submission³⁵ and hearing appearance signal some areas where we would be very willing to work with governments on new regulation, including in relation to hate speech.

Protecting the safety and security of Australians online is a continuous task. It is a responsibility of multiple stakeholders, including governments, industry and law enforcement. No one organisation can rest on their laurels: we all need to continuously consider what more we can do. We welcome all opportunities to deepen our collaboration with important government departments such as Home Affairs.

We recognise that it is a major undertaking for government departments to maintain an accurate and up-to-date understanding of different tech companies' approaches to these issues, including as departments are restructured or staff move on. We want to continue to work closely with the Department, Ministers and officials to improve the understanding of Meta's work.

³⁵ Meta, 'Submission to the Select Committee on Social Media and Online Safety - submission 49', *Parliament of Australia*, https://www.aph.gov.au/Parliamentary_Business/Committees/House/Social_Media_and_Online_Safety/SocialMediaandSafety/Submissions

Question 19

Facebook operates a Dangerous Organisations and Individuals list to help prevent highly risky groups from organising violence on your platform – is that correct?

How many Australian organisations are on this list?

Meta response:

Meta's Community Standards prohibit any organisation or individual that proclaims a violent mission or is engaged in violence from having a presence on Meta's platforms.³⁶ Specifically, we do not allow on our platform dangerous organisations and individuals, including:

- terrorist organisations and terrorists
- hate organisations, and their leaders and prominent members,
- criminal organisations,
- mass / multiple murderers (including attempted murderers).

As well as removing these groups, we do not allow content that praises, supports or represents these groups.

Although our enforcement will not always be perfect, we have made significant progress in detecting and removing terrorist and extremist groups on our services. We have banned more than 270 white supremacist organisations globally and we have removed about 1,000 militarised social movements from our platform.³⁷ Some of the individuals and organisations designated in Australia include Blair Cottrell, Neil Erickson, Tom Sewell, the Lads Society, the United Patriots Front, True Blue Crew and the Antipodean Resistance.

This is an adversarial space. We do not make our full list of dangerous organisations or individuals publicly available, due to the potential safety and security ramifications, as well to avoid bad actors evading our enforcement. Country-specific numbers of dangerous organisations or individuals are challenging to prepare and not necessarily meaningful for policymakers, because (1) some dangerous organisations span multiple

³⁶ See Meta, *Community Standards - Dangerous individuals and organisations*, <https://www.facebook.com/communitystandards/>

³⁷ Meta, 'An update to how we address movements and organizations tied to violence', *Meta Newsroom*, blog post updated 19 January 2021, <https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/>

countries; and (2) Australians may be negatively impacted by the online activities of a dangerous organisation or individual, even if they are not based in Australia.

We would be very happy to arrange a confidential briefing for the Committee on our list, if desired.

Question 20

How many Facebook staff either working in Australia or with work experience in Australia are dedicated to identifying Australian organisations for this list

Meta response:

There are more than 350 people at Meta whose primary job is countering terrorist and violent content on our platforms. This team includes former academics who are experts on counterterrorism, former prosecutors and law enforcement agents, investigators and analysts, and engineers. Within this specialist team, they speak nearly 30 languages. This number does not include other staff within the organisation who play an important role in combatting terrorism and organised hate but may also focus on other issues, like content moderation or public policy teams.

The role of the Dangerous Individuals and Organisations team is to:

- Study new trends in speech and adversarial behavior related to violence
- Continuously monitor the evolution of organisations engaging in this behaviour
- Develop Meta's strategy in response to this challenge

They also partner with a range of external experts globally in terrorism, violent extremism, cyber intelligence and online behaviour; and commission independent research to better tailor our response.

We have a team focused on countering terrorism and violent extremism in the Asia Pacific. They have dedicated their entire careers to this work and, outside of their time at Meta, are considered industry leaders in this space. They have a deep understanding of the terrorism and organised hate landscape across the entire region - including in Australia.

The Dangerous Individuals and Organisations team is part of a broader team of 40,000 people at Meta, who are focused on safety and security. This includes dedicated teams focused on Australia.

Question 21

Your submission mentions an “*Australia specific Combatting Online Hate Advisory Group*” – who is on this? What is their feedback about Facebook’s performance on this issue?

Meta response:

We established an Australia-specific Combatting Online Hate Advisory Group in October 2020, which consists of 15 representatives. The following organisations have consented to us disclosing their membership of this Group in response to this Question on Notice, within the timeframe requested by the Committee:

- Andre Oboler - Online Hate Prevention Institute
- Kosta Lucas
- Priscilla Brice
- Teddy Cook - ACON.

Some members have specifically asked not to be publicly named, especially given the potential risk to themselves in being seen to be discussing these issues. We can confidentially brief the Committee further in camera, if desired.

It also does not fully capture the range of Australian academics, experts and representatives of community groups with whom we have consulted on the issue of hate speech, only those who have agreed to be part of this smaller and more intensive working group.

We have heard feedback from the group in a number of respects. Some of the key lessons that we have taken away to date include:

- The need for platforms to better understand the trans experience.
- The need for a more holistic understanding of harmful conspiracy theories.
- The need for more research into specific aspects of online hate.
- A view from the group that additional investment from Meta in counterspeech initiatives should not be the priority at this time.

We have taken a number of steps directly in response to this feedback, including:

- Arranging for workshops for all Australian staff - with mandatory attendance by executives - on encouraging trans-affirming workplaces; and commissioning research on trans experiences to inform our Community Standards.

- Incorporating feedback on harmful conspiracy theories into our global policy development process.
- Commissioning research on anti-Asian hate online and offline, specifically since the beginning of the COVID-19 pandemic.
- We have directed investment into research rather than counterspeech programs, on the basis of advice from the advisory group.

Question 22

Your submission highlights that over the past three years the Morrison government has initiated 18 major government or parliamentary inquiries or consultations impacting digital platforms.

You've warned that

"Policy makers should be alive to the risk of overlapping, duplicative or inconsistent rules across different laws. Indeed, many of the online safety-related laws and regulations that have already been passed by Parliament have yet to be implemented. Policy makers will be able to develop more effective regulation if there is consideration given to properly understanding the effectiveness of existing regulation first."

What are the risks to good policy outcomes of this government operating parallel, overlapping policy development processes?

Meta response:

There can be a range of risks to good policy outcomes resulting from duplicative or inconsistent regulation. As the OECD points out:

"Rules and procedures that do not correspond to genuine risks tend to result in higher costs and burdens, without providing real benefits. Those that do not effectively target the causes of risks, based on findings from research and evidence, likewise fail to deliver. Regulation that is not useful or effective decreases public trust, and harms the economy."³⁸

More specifically, we would elaborate on four risks in particular, including:

- **Duplicative or inconsistent regulation can lead to confusion.** It's in the interests of Australian consumers for rules to be clear and consistently applied. Laws need to be well-understood by industry, regulators and policymakers, not-for-profit organisations, academics, other experts - and the Australian community at large.
- **Duplicative or overlapping laws will be less effective if they do not account for assessments of whether existing laws are working.** For example, the Online Safety Act has only come into effect in January 2022. To advance further online safety laws without allowing sufficient time for this legislation to operate (and without a

³⁸ OECD, *OECD Regulatory Policy Outlook 2021*,
<https://www.oecd.org/gov/regulatory-policy/oecd-regulatory-policy-outlook-2021-38b0fdb1-en.htm>

comprehensive evaluation), risks governments pursuing new laws that do not learn the lessons of previous reform.

- **Unnecessary regulations can have unintended consequences and economic impacts.** If new regulations are not fit-for-purpose and targeted, they risk unintended consequences - such as chilling innovation, discouraging foreign investment, inhibiting the ability for companies to launch new products for Australians, making it harder for start-ups and small businesses to enter the market, or limiting the ability of Australians to express themselves and fully participate in the online world.
- **Consistency in regulations between like-minded countries helps to reduce the risks associated with internet fragmentation.**

Question 23

When asked at recent public hearings whether certain statements would breach a platform's terms of service and allow for removal, a number of social media companies outlined that the answer depended on the context of the comments.

Below are some examples of abusive and derogatory comments posted online. Can you outline to the committee the context in which these statements would not breach your policies and remain on your site?

Would the context differ for a private individual and a public figure?

- Language suggesting a woman should put a bag over her head to suffocate herself
- "Take half a brick and hide behind a bush"
- "I bet she rages so hard a natural disaster occurs every time she has her period"
- "That woman is the reason nature designed the human hand to grasp a penis in a pleasing manner"
- "Cavorting whore"
- "Every homophobic whore deserves to have their c*nt sliced open with a chainsaw.

Including the filthy sluts in Australian Parliament... I'm not advocating violence, but if any of those bigots was set on fire, I'd toast marshmallows in the flames."

Meta response:

Of the six comments provided by the Committee in this Question on Notice, five would violate our policies regardless of whether they are directed at a private individual or public figure (based on the information available to us). As outlined in our submission and during our appearance before the Committee, we have steadily updated our policies over the years to provide greater protections to public figures to protect them against mass harassment and degrading or sexualised attacks on our services.³⁹ These policy updates are in addition to the new tools that we have developed to support public figures such as controlling comments, hidden words and "Limits".⁴⁰

We have responded below in response to each comment. In terms of context, we have assumed that, for public figures, the comment is posted on their Page or otherwise directed at them.

³⁹ See our latest announcement:

<https://about.fb.com/news/2021/10/advancing-online-bullying-harassment-policies/>

⁴⁰ Please see pages 24–27 of Meta's submission to this Inquiry for more details about these tools.

- Language suggesting a woman should put a bag over her head to suffocate herself: Any calls to commit self-injury or suicide would violate our policies and would be removed regardless of whether it was aimed at a private individual or a public figure.
- “Take half a brick and hide behind a bush”: If we had additional context that this comment was intended to contain a veiled threat of violence, then this would violate our policies and be removed regardless of whether it was aimed at a public figure or a private individual.
- “I bet she rages so hard a natural disaster occurs every time she has her period”: This does not violate our policies.
- “That woman is the reason nature designed the human hand to grasp a penis in a pleasing manner”: This comment would violate our policies and would be removed regardless of whether it was aimed at a private individual or a public figure.
- “Cavorting whore”: This comment would violate our policies and would be removed regardless of whether it was aimed at a private individual or a public figure.
- “Every homophobic whore deserves to have their c*nt sliced open with a chainsaw. Including the filthy sluts in Australian Parliament... I’m not advocating violence, but if any of those bigots was set on fire, I’d toast marshmallows in the flames”: This comment would violate our policies and would be removed regardless of whether it was aimed at a private individual or a public figure.