



Meta's submission to the Inquiry into the Administration of the Referendum into an Aboriginal and Torres Strait Islander Voice

MAY 2023

Executive summary

Meta welcomes the opportunity to participate in the inquiry being conducted by the Senate Finance and Public Administration References Committee into the administration of the referendum into an Aboriginal and Torres Strait Islander Voice.

The referendum will be a significant moment for Australia. Many Australians will use digital platforms to engage in advocacy, express their views, or participate in democratic debate.

Meta is committed to playing our part to safeguard the integrity of electoral events, including the referendum. We've been involved in more than 200 elections around the world since 2017, and at a global level, we now have more than 40,000 people working on safety and security at Meta. We've invested more than \$16 billion (~AU\$23 billion) in teams and technology to enhance safety and security since 2016.

Our priorities are to protect people's voice on our services, help them participate in the civic process, and combat potential risks that can arise during election campaigns, such as foreign interference or sharing of misinformation. These risks pose a critical, continuous challenge for governments, industry, media, civil society and academia. Cross-sector cooperation is essential to address them.

Meta will announce the measures we are taking specifically for the referendum later in 2023, following the conclusion of our consultation with relevant stakeholders and Meta's Aboriginal and Torres Strait Islander Advisory Group. In the meantime, to assist the Committee, this submission provides information about the approach Meta takes to safety and security which is in place at all times, and provides the foundation of our approach to the referendum.

While the Committee's inquiry is focussed on the administration of the referendum, Meta remains strongly committed to working with First Nations communities and our submission also provides background on our work in this area.

Our submission provides information on our Community Standards (which outline what is and is not allowed on Facebook and Instagram), how we enforce our policies, and our industry-leading transparency requirements for ads about social issues, elections and politics.

In an effort to assist the Committee with an understanding of our work to protect election integrity, we have included details of our efforts during the 2022 Australian federal election. The impact of our efforts in that campaign is clear. During the election campaign:

- We took action on over 25,000 pieces of content across Facebook & Instagram for violating our Harmful Health Misinformation policies.
- We displayed warnings on over 3 million distinct pieces of content on Facebook (including reshares) based on articles written by our third party fact checking partners.
- We took action on over 91,000 pieces of content on Facebook and over 40,000 pieces of content on Instagram in Australia for violating our hate speech policies.
- We took action on over 200,000 pieces of content on Facebook and over 46,000 of content on Instagram in Australia for violating our Community Standards on violence and incitement.
- We rejected around 17,000 ads for not complying with our political and social issue ads enforcement policies.
- We also worked with the Australian Electoral Commission (AEC) to release two prompts to encourage people to vote, and direct users to the AEC website. These prompts were seen by over 23 million Australians, and almost 11 million Australians, respectively.

Meta will continue to be a constructive partner to the Government in the administration of the referendum. In line with our approach to the 2022 federal election, we will conduct a series of briefings with key stakeholders - including the AEC, the Australian government, parties and advocacy groups - on our approach to integrity across our platforms. We welcome the opportunity to engage with Australian policymakers on the referendum further.

Table of contents

Meta’s work with First Nations communities	5
Meta’s approach to electoral events	10
Policies	10
Hate speech	11
Violence and incitement	12
Fake accounts	12
Misinformation	12
Coordinated inauthentic behaviour	19
Enforcement	20
Hate speech	22
Violence and incitement	23
Fake accounts	24
Misinformation	25
Coordinated inauthentic behaviour	25
Partnerships	28
Transparency and accountability	32

Meta's work with First Nations communities

Meta is committed to working with First Nations peoples as they use our platforms to connect to their communities, share their stories, celebrate language and culture, or grow their business. We work to achieve this in line with the following principles: Relationships & Respect, Opportunities and Governance.

Relationship & respect

Meta is committed to deepening cultural capacity and understanding about First Nations culture and history across our organisation, and seeking feedback on how we can integrate these learnings into our products and programs. Meta is a sponsor of NAIDOC week,¹ and were proud to sponsor and participate in the Garma Festival in 2022 and also, participate in 2019 as part of deepening our cultural capacity within the organisation.

Our work is informed by an Aboriginal and Torres Strait Islander Advisory Group which was established in 2021 and comprises Aboriginal and Torres Strait Island academics, community leaders, small business owners and educators. The Group typically meets quarterly to provide feedback on Meta's policies, products and programs, and inform Meta's approach to building a safe and positive experience for Aboriginal and Torres Strait Islander peoples across Meta's platforms.

Four examples of ways in which we have worked to integrate the representation of First Nations people in our products include:

- Connect to Country campaign. In 2021, Meta partnered with Campfire x and more than 10 Aboriginal organisations and land councils to launch a campaign asking: "Where you from?". The campaign was launched during NAIDOC week and encouraged Australians to acknowledge Country and learn more about First Nations communities. Meta also worked with Campfire x to launch a pilot series of unique First Nations stories that was shared on Facebook and Instagram to people who live in corresponding areas, connecting them with the land they're on.²

¹ A Sloane, 'Elevating and celebrating First Nations people this NAIDOC Week 2022', *Meta Australia Blog*, 3 July 2022, <https://medium.com/meta-australia-policy-blog/elevating-and-celebrating-first-nations-people-this-naidoc-week-2022-eadc5cea7e7b>

² A Sloane, 'Connect to Country and ask 'Where you from?', *Meta Australia blog*, 2 July 2021, <https://medium.com/meta-australia-policy-blog/connect-to-country-and-ask-where-you-from-60f64283f0c9>

Connect to Country 'Where you from?' campaign



- The 'Deadly' augmented reality effect. In 2022, Meta partnered with Indigitek, a non-profit organisation that supports the participation of Indigenous Australians in the tech industry, and Awesome Black, a First Nations-led social enterprise supporting Indigenous creators, to create a new augmented reality (AR) effect for Instagram. The AR effect was designed by First Nations digital artist, Rubii Red, and brings to life Indigitek's 'Deadly' emote from their Twitch channel. The Deadly AR pin was shared at the Pinny Arcade Expo Australia in Australia in October 2022.³
- Diverse Voice campaign for NAIDOC week 2022. In 2022, Meta launched the Diverse Voices campaign in Australia to enable small businesses from First Nations communities to find success in their shift to digital with Meta technologies. The campaign featured five Indigenous entrepreneurs, including the Brolga Dance Company.

³ J Machin, 'Indigitek, Awesome Black and Meta launch new AR 'Deadly' emote filter for PAX 2022', *Meta Australia Blog*, 7 October 2022, <https://medium.com/meta-australia-policy-blog/indigitek-awesome-black-and-meta-launch-new-ar-deadly-emote-filter-for-pax-2022-447f289411bb>

- The #SeeTheIndigenousWorldThruMyEyes initiative on Instagram. This global program, launched in 2023, showcased Indigenous creators in Australia, Canada, and the US. The program aims to highlight Indigenous cultures and businesses through Indigenous Creators, using Ray-Ban Stories. In Australia, the partnership was launched with First Nations consultancy Campfire x and EssilorLuxottica, and shares the culture and experiences of three local Indigenous creators - Summer Simon, Felicia Foxx and KurenMusic - to Instagram followers across Australia and around the world.⁴

Opportunities

Our work is focussed on developing and maintaining mutually beneficial relationships with First Nations peoples, communities and organisations. Recently, we have worked towards opportunities via research and programs.

Meta continues to invest in research to better understand First Nations experiences with social media. For example:

- In 2020, we commissioned Professor Bronwyn Carlson from Macquarie University to conduct research on Indigenous women and LBGTQI+ people and violence on Facebook.⁵
- In 2020, we commissioned Professor Tristan Kennedy from Macquarie University (and now Monash University) to conduct research on Indigenous peoples' experiences with harmful content online.⁶ The research found that while Indigenous communities disproportionately experience harmful content online, social media also provides the opportunity for Indigenous peoples to express themselves, empowering communities to share stories and speak their truth.

⁴ Instagram launches campaign 'see the indigenous world through my eyes', Bandt, 24 March 2023, <https://www.bandt.com.au/instagram-launches-campaign-see-the-indigenous-world-through-my-eyes/>

⁵ Meta Research, 'Announcing the winners of Facebook's request for proposals on misinformation and polarisation', *Meta Research*, 7 August 2020, <https://research.facebook.com/blog/2020/8/announcing-the-winners-of-facebooks-request-for-proposals-on-misinformation-and-polarization/>

⁶ T Kennedy, 'Indigenous peoples' experiences of harmful content on social media', *Macquarie University*, 2020, https://research-management.mq.edu.au/ws/portalfiles/portal/135775224/MQU_HarmfulContentonSocialMedia_report_201202.pdf

- In 2021, Meta also supported the Australian Media Literacy Alliance's first media literacy survey and the report 'Towards a National Strategy for Media Literacy'. The report demonstrated that there is more work required to understand the media literacy of First Nations Australians.⁷ For this reason, Meta has partnered with Professor Tristan Kennedy to conduct research on the cultural nuances of misinformation and the impacts on Indigenous Australian communities, so that Meta can best tailor our programs to this area going forward. This research is currently in development.
- In 2022, we commissioned Jasper Garay from the University of Sydney to conduct research which will explore perceptions of augmented reality (AR) and virtual reality (VR) within First Nations communities, and identify ways in which AR and VR can increase cultural promotion and wellness for first nations people.⁸

Meta also engages in programs to directly support First Nations communities and small businesses, as we know that First Nations creators and small businesses use Meta's platforms to engage broad audiences and build their brands. For example:

- In 2019, we worked with the Alannah and Madeline Foundation and the Stars Foundation to create a new program, Safe Sistas,⁹ which supports the online safety of young Indigenous women to respond to the issue of non-consensually shared intimate images in a culturally relevant and safe way. The program reached 857 young girls in Years 7 - 12 in remote and regional communities across the Northern Territory, Queensland and Victoria.
- In 2022, Meta worked with First Nations owned Trading Blak and Nungala Creative to launch the #BuyBlak campaign.¹⁰ The BuyBlak campaign launched in line with Indigenous Business Month and ahead of Black Friday, and is a celebration of First Nations business excellence. The campaign promoted a series

⁷ Australian Media Literacy Alliance, 'Towards a National Strategy for Media Literacy', *AMLA*, 25 October 2021, https://medialiteracy.org.au/wp-content/uploads/2021/10/AMLA-Consultation-Workshop-Report_UPDATE-25-10-2021.pdf

⁸ Meta Research, 'Announcing the winners of the 2022 Meta AR / VR policy research request for proposals for the Asia Pacific region', *Meta Research*, 30 September 2022, <https://research.facebook.com/blog/2022/9/announcing-the-winners-of-the-2022-meta-arvr-policy-research-request-for-proposals-for-the-asia-pacific-region/>

⁹ Alannah & Madeline Foundation, *Helping Sistas be safer*, <https://researchers.mq.edu.au/en/publications/safe-sistas-evaluation-report>

¹⁰ A Sloane, 'Meta partners with Trading Blak to equip First Nations businesses with the right online sales tools to get more people to #BuyBlak', *Meta Policy Blog*, 29 September 2022, <https://medium.com/meta-australia-policy-blog/meta-partners-with-trading-blak-to-equip-first-nations-businesses-with-the-right-online-sales-tools-e7a43f667e95>

of assets and short videos educating businesses on how to use Meta's tools to reach more Australians across Meta's platforms. It also encouraged Australians to shop with these businesses.

- In 2022, Meta partnered with Screen Australia to launch the First Nations Creator Fund for the second year in a row.¹¹ The Fund aims to foster First Nations talent, amplify their voices online and create long term career prospects in social media and content creation. The Fund provided 10 Aboriginal and Torres Strait Islander social media creators with access to an immersive program including in-person training, practical workshops, mentoring, career connections, equipment and content funding.
- In 2022, Meta supported Indigital, Australia's first Indigenous edu-tech company, to conduct a series of workshops with First Nations communities, to educate on immersive technologies, and support them to create immersive experiences which express their knowledge and stories in a way that respects the knowledge, language and lore.

Governance

In 2019, Meta established a First Nations Employee Committee which oversees the planning and implementation of Meta's First Nations program of work and managing Meta's involvement in cultural moments.

¹¹ Screen Australia, 'Instagram and Screen Australia announce recipients of 2022 First Nations Creator Program', *Screen Australia*, 14 October 2022, <https://www.screenaustralia.gov.au/sa/media-centre/news/2022/10-14-first-nations-creator-program-22>

Meta's approach to electoral events

Meta is committed to investing in protecting elections online - not just during election periods but at all times. Since 2016, we have made large investments in safety and security with more than 40,000 people working on these issues.

Meta has developed a comprehensive approach to safeguarding the integrity of electoral events on our platforms. This work falls into five areas, and continues to be updated to reflect feedback from experts, governments and the community:

1. **Policies** that set out what type of content, actors and behaviour is and is not allowed on our services;
2. Processes to **enforce** our policies;
3. **Partnerships** with local stakeholders to ensure our strategy and the measures we take to inform our users about the election process have the most effective impact; and
4. **Transparency and accountability** about organic content and advertising on political and social issues, and the decisions we take relating to content on our services.

These areas are each outlined in further detail below.

Policies

First, our policies, known as our Community Standards,¹² outline what is and is not allowed on Meta's services. These policies are developed based on a range of values to help combat abuse including safety, privacy, authenticity, voice, and dignity.¹³

Our priority is to provide a platform that allows users to discuss topics and express their opinions, whilst limiting the spread of content that may be harmful.

Our policies are informed by feedback from our community, and the advice of experts in fields such as technology, public safety, child safety and human rights. To ensure that everyone's voice is valued, we take great care to craft policies that are inclusive of

¹² See Meta, *Community Standards*, <https://www.facebook.com/communitystandards>

¹³ Monika Bickert, *Updating the values that inform our community standards*, <https://about.fb.com/news/2019/09/updated-the-values-that-inform-our-community-standards/>

different views and beliefs, in particular those of people and communities that might otherwise be overlooked or marginalised.

Some of our policies may be especially relevant to electoral events, such as:

- Hate speech
- Violence and incitement
- Fake accounts
- Misinformation
- Coordinated inauthentic behaviour.

These are outlined in further detail below.

In the lead up to an election or referendum, Meta will conduct a series of briefings with key policy stakeholders on our policies, and the process for referring and reviewing content during the referendum campaign.

Hate speech

Hate speech is always harmful, but in the context of an electoral campaign it can create an environment of intimidation and exclusion at a time when people want to use their voice to express their opinion and connect with their communities.

We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That is why we don't allow hate speech on our services, as it creates an environment of intimidation and exclusion, and in some cases may promote offline violence.

We define hate speech as a direct attack against people on the basis of what we call protected characteristics which include race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.¹⁴

We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation. This goes well beyond what is required in Australian legislation.

¹⁴ Meta, *Community Standards - Hate Speech*, <https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/>

We also prohibit content that describes or negatively targets people with slurs, where slurs are defined as words that are inherently offensive and used as insulting labels for the above characteristics.

Violence and incitement

We aim to prevent potential offline harm that may be related to content on our services.¹⁵ We remove content, disable accounts and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety.

While we understand that people commonly express disagreement during an electoral campaign by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence.

Fake accounts

We consider authentic communications to be a central part of people's experience on Facebook.¹⁶ People find value in connecting with their friends, family, and issues they care about, and we want them to be able to trust the people they interact with. For this reason, authenticity has long been a requirement of our Community Standards.

Fake accounts can often be the vehicle for harmful content, including misinformation. Our goal is to remove as many fake accounts as we can. These include accounts created with malicious intent to violate our policies and personal profiles created to represent a business, organisation or non-human entity, such as a pet.

Misinformation

Meta takes a global approach to combatting misinformation, and we constantly update our efforts in response to feedback, research, and changes in the nature of misinformation.

¹⁵ Meta - *Community Standards - Violence and incitement*, <https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/>

¹⁶ Meta, *Community Standards - Misrepresentation*, <https://www.facebook.com/communitystandards/misrepresentation/>

Before turning towards Meta's approach to these issues, it is important to recognise the distinctions between 'misinformation' and 'disinformation', because the policy concerns underlying each differ, and the most appropriate responses from platforms like Meta will also be different.

Misinformation is often used interchangeably with disinformation. Disinformation refers to false information that is shared intentionally to mislead others, therefore Meta's approach to this issue focuses on the *actor* and *behaviour*.

Misinformation refers to *content* that is misleading or false, but may be shared without an intent to mislead. While there may be some overlap (actors engaged in disinformation may also utilise misinformation), disinformation and misinformation are not the same issue.

The closest term that Meta uses to describe disinformation is coordinated inauthentic behaviour (CIB). We have described our approach to CIB in the following section.

Our policies on misinformation outline that we will:

- **Remove** misinformation that could directly contribute to the risk of imminent physical harm, interference with the functioning of political processes, and certain highly deceptive manipulated media; and
- **Reduce** the spread of misinformation that is identified and verified as false by independent third party fact-checkers.

We are committed to working with policymakers and partners around the world to combat misinformation. In particular, in 2020 we became a founding member and signatory to the Australian Disinformation and Misinformation Industry Code.¹⁷ The Code is a major step in establishing a regulatory framework around industry's work to combat misinformation and disinformation, with other countries around the world looking to emulate this approach.

Since the launch of the Code, Meta has publicly released two transparency reports outlining our specific commitments to meet the obligations outlined in the voluntary code. We will report again on these efforts in May 2023.¹⁸

¹⁷ J Machin, 'Facebook's response to Australia's disinformation and misinformation industry code', *Meta Australia Blog*, 21 May 2021, <https://medium.com/meta-australia-policy-blog/metas-annual-transparency-report-on-australia-s-disinformation-and-misinformation-industry-code-8023640f8de4>

¹⁸ You can find our two transparency reports here <https://diqi.org.au/disinformation-code/transparency/>

Remove

Meta's priority is to provide a platform that allows users to discuss topics and express their opinions, whilst limiting the spread of misinformation that may be harmful. We remove the following types of misinformation:

- **Misinformation that is likely to directly contribute to a risk of interference with people's ability to participate in an election or census process.** We continue to update our election integrity policies to reflect the realities of current elections. In recent years, we broadened our policies to expressly ban election-related misinformation that may constitute voter fraud and / or interference.¹⁹ We work to proactively identify and remove this type of harmful content, even if it comes from a politician.²⁰

Under our policies, we prohibit:

- misrepresentation of the dates, locations, times, and methods of voting or voter registration (for example: false claims that you can vote using an online app).
- misrepresentation of who can vote, qualifications for voting, whether a vote will be counted, and what information or materials must be provided in order to vote.
- Misinformation about whether a candidate is running or not.
- Explicit false claims that people will be infected by COVID-19 (or another communicable disease) if they participate in the voting process.

These policies apply to all elections, including referendums.

- **Misinformation that is likely to directly contribute to imminent physical harm.** We have had a policy on Misinformation and Harm since 2018, which has been used in instances such as when we removed harmful health misinformation during the measles outbreak in Samoa towards the end of 2019.

As soon as the COVID-19 pandemic started, we started working with experts around the world - in particular, the World Health Organization - to identify COVID-related claims that could contribute to imminent physical harm.

¹⁹ J Leinwand, 'Expanding our policies on voter suppression', *Meta Newsroom*, 15 October 2018, <https://about.fb.com/news/2018/10/voter-suppression-policies/>

²⁰ Meta, *Community Standards - Misinformation*, <https://transparency.fb.com/en-gb/policies/community-standards/misinformation/> and; Meta, *Community Standards - Coordinating and publicizing crime*, <https://transparency.fb.com/en-gb/policies/community-standards/coordinating-harm-publicizing-crime/>

In December 2020, we expanded our policy to cover false claims about COVID vaccines, and in January 2021 we started applying the policy to claims about vaccines generally.²¹

- **Material that violates our Violence-Inducing Conspiracy Network policy.** Our Community Standards state that we will remove conspiracy theory material that advocates for violence and “violence-inducing conspiracy theories”. This specifically includes any Facebook Pages and Instagram accounts associated with QAnon, which have been known to spread misinformation during election campaigns in the past.

As of August 15, 2022, we have identified over 1,151 militarised social movements to date and in total, removed about 4,200 Pages, 20,800 groups, 200 events, 59,800 Facebook profiles and 8,900 Instagram accounts. We’ve also removed about 4,200 Pages, 12,000 groups, 840 events, 67,200 Facebook profiles and 38,800 Instagram accounts for violating our policy against QAnon.²²

We are continuing with consultation and consideration of a potential harmful conspiracy theory policy that accounts for harms broader than violence, as advocated by QAnon.

- **Manipulated videos, also known as “deepfakes”, that violates our Manipulated Media policy.** After consulting with more than 50 global experts with technical, policy, media, legal, civic and academic backgrounds, we announced in 2020 that we will be removing manipulated videos if: (1) it has been edited or synthesised – beyond adjustments for clarity or quality – in ways that aren’t apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say; and (2) it is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic.²³

²¹ K Jin, ‘Keeping people safe and informed about the coronavirus’, *Meta Newsroom*, 18 December 2020, <https://about.fb.com/news/2020/12/coronavirus/#removing-covid-vaccine-misinformation>

²² 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/>.

Note - Meta began reporting our enforcement data for the Violence-Inducing Conspiracy Network policy in August 2020. Only global data is available, as this includes Groups, Pages and Events which can compromise users based in numerous different countries.

²³ M Bickert, ‘Enforcing Against Manipulated Media’, *Meta Newsroom*, 6 January 2020, <https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/>.

Reduce

For content that does not violate our Community Standards but is still false, problematic or low-quality, we reduce its distribution.

Our approach to content distribution, outlined in our Content Distribution Guidelines, was developed in consultation with more than 100 stakeholders who are experts on how to limit the spread of problematic content and bring more transparency to our efforts.²⁴

The Guidelines outline the steps we take to limit the spread of problematic content, including in a number of areas that relate to misinformation and disinformation, such as:

- Content that has had its distribution reduced because it's been found to be false by fact-checkers;
- Sensational health content and commercial health posts;
- Posts from broadly untrusted news publishers; and
- Behaviour that represents artificially boosting engagement or views.

We take a number of steps to identify content which should have its distribution reduced:

- **Third party fact checking program.** We have commercial agreements with independent third-party fact-checking organisations to review and rate the accuracy of posts on Facebook and Instagram. Since 2016, Meta has contributed more than \$100 million globally to support our fact-checking efforts.²⁵

In March 2022, ahead of the Australian federal election, we announced the expansion of our third-party fact-checking program in Australia to include RMIT FactLab, alongside Australian Associated Press and Agence France Presse.²⁶ All fact-checks by these partners are publicly available on their websites.²⁷

²⁴ 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/>

²⁵ C Alexander, 'Facebook launches accelerator challenge for global fact-checkers to expand reach of reliable information', *Meta Journalism Project*, 26 August 2021,

<https://www.facebook.com/journalismproject/accelerator-fact-checkers>

²⁶ J Machin, 'How Meta is preparing for the 2022 Australian election', *Meta Australia blog*, 15 March 2022, <https://medium.com/meta-australia-policy-blog/how-meta-is-preparing-for-the-2022-australian-election-c627e6b3c0a8>

²⁷ Agence France Presse Australia, Fact Check, <https://factcheck.afp.com/afp-australia>; Australian Associated Press; AAP Fact Check, <https://www.aap.com.au/category/factcheck>; RMIT FactLab <https://www.rmit.edu.au/about/schools-colleges/media-and-communication/industry/factlab/debunking-misinformation>

In the lead up to the Australian election, we also took a number of additional steps to enhance Australia's fact-checking capability and capacity, including:²⁸

- Providing one-off grants to each of our fact-checkers to increase their capacity in the lead up to the election;
- Working with First Draft (now RMIT FactLab), the global misinformation and media literacy organisation, to increase misinformation monitoring and analyses; and
- Providing pre-election training for Australian journalists on how to prevent amplifying mis and dis information.

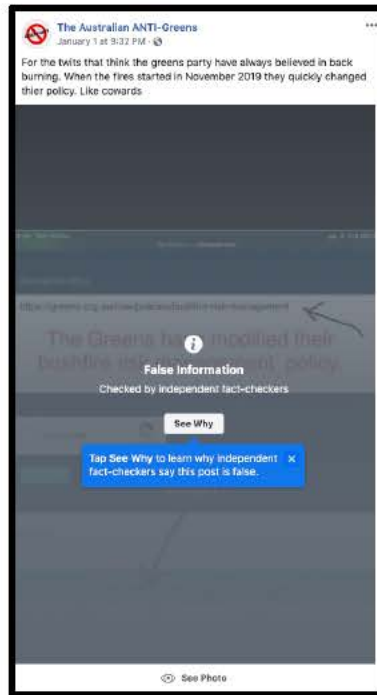
It's important to note that Australians also benefit from Meta's international approach to fact-checking. As of December 2022, Meta partners with over 90 fact checking partners covering more than 60 languages.²⁹ An Australian user will see a warning label on content that has been fact-checked by an international fact-checking partner. Content found to be false by our international fact-checking partners will be demoted in an Australian user's Feed, meaning there is less chance of them seeing it.

- **Warning labels.** Once a third-party fact-checking partner rates a post as 'false', we apply a warning label that indicates it is false and shows a debunking article from the fact-checker. It is not possible to see the content without clicking past the warning label.

²⁸ J Machin, 'How Meta is preparing for the 2022 Australian election', *Meta Australia blog*, 15 March 2022, <https://medium.com/meta-australia-policy-blog/how-meta-is-preparing-for-the-2022-australian-election-c627e6b3c0a8>

²⁹ Meta, 'Where we have fact checking', *Meta for Media*, <https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking/partner-map>

Warning label on a fact-checked piece of content



Based on one fact-check, we're able to kick-off similarity detection methods that identify duplicates of debunked stories. Using this technology, we are able to limit the distribution of similar posts: for example between April and June 2022, we displayed warnings on over 200 million distinct pieces of content on Facebook (including re-shares) globally based on over 130,000 debunking articles written by our fact-checking partners.³⁰

- **Reducing the spread of misinformation in private messages, Pages, groups accounts or websites.** Pages, Groups, accounts, and websites that repeatedly share content that is marked false by fact-checkers will see reduced distribution of all their posts. Since the beginning of the pandemic, we've removed over 3,000 accounts, Pages, and groups for repeatedly violating our rules against spreading COVID-19 and vaccine misinformation.³¹

³⁰ M Bickert, 'Community Standards Enforcement Report, Second Quarter 2022, *Meta Newsroom*, 25 August 2022, <https://about.fb.com/news/2022/08/community-standards-enforcement-report-q2-2022/>

³¹ M Bickert, 'How we're taking action against vaccine misinformation superspreaders, *Meta Newsroom*, 18 August 2021, <https://about.fb.com/news/2021/08/taking-action-against-vaccine-misinformation-superspreaders/>

- **Limiting the ability to forward material via private messaging.** We have instituted strict forward limits for WhatsApp and Messenger to introduce friction in the experience and constrain message virality.³² On WhatsApp, we have placed a limit on “frequently forwarded” messages - defined as a message that has already been forwarded on five times - so they can only be sent to one chat at a time. After implementing this change, WhatsApp saw a 70 per cent reduction in the number of highly forwarded messages sent on WhatsApp.³³

Coordinated inauthentic behaviour

The closest term to foreign interference or disinformation that Meta uses is coordinated inauthentic behaviour (CIB). Both foreign interference and CIB rely on two elements: inauthenticity - where users misrepresent themselves, through fake profiles or non-transparent behaviours, and coordination - where groups of accounts work together with the intention to deceive users.

Specifically, our policies prohibit people engaging in inauthentic behaviour, which, as mentioned above, includes creating, managing, or otherwise perpetuating accounts that are fake, and accounts that have fake names.

We also prohibit accounts that participate in coordinated inauthentic behaviour (CIB),³⁴ which is commonly referred to as disinformation. We define CIB as “any coordinated network of accounts, Pages and Groups that centrally relies on fake accounts to mislead Meta and people using our services about who is behind it and what they’re doing”.³⁵

CIB, as we define it, will be slightly broader than the strict interpretation of “foreign interference”, as CIB may include inauthentic coordination by domestic actors, and it may include CIB that is financially motivated (eg, scams) rather than politically motivated. We take action on CIB according to the behaviour of the actors in the network, not the content they post.

³² WhatsApp, About forwarding limits, <https://faq.whatsapp.com/general/chats/about-forwarding-limits>; J Sullivan, ‘Introducing a forwarding limit on Messenger’, *Meta Newsroom*, 3 September 2020, <https://about.fb.com/news/2020/09/introducing-a-forwarding-limit-on-messenger/>

³³ WhatsApp, ‘About forwarding limits’, *Help Centre*, <https://faq.whatsapp.com/1053543185312573>

³⁴ Meta, *Coordinated Inauthentic Behavior Explained*, <https://newsroom.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/>

³⁵ Meta, ‘Threat Report: The State of Influence Operations 2017-2020’, *Meta Newsroom*, May 2021, <https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf>

Enforcement

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

Meta employs a number of different measures to enforce our policies, we:

- **Proactively detect and remove violating content or suspicious behaviour.** Meta uses a combination of technology and human reviewers to proactively detect and remove the vast majority of violating content on the platform before it's reported.

We have scaled our enforcement to review millions of pieces of content across the world every day, and use our technology to help detect and prioritise content that needs review.

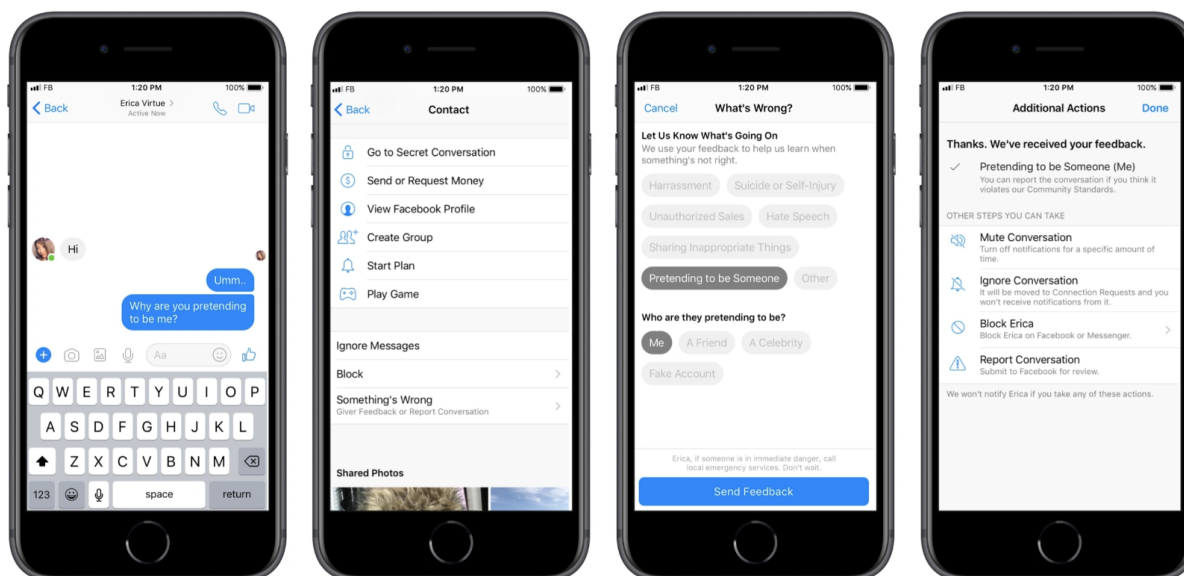
We've made considerable progress in our detection over the last few years: for many categories of seriously harmful content, we are proactively detecting more than 95 per cent of violating content ourselves, before a user needs to report it to us.

- **Develop channels for individual users or onboarded government agencies to report content and ads that may contravene our policies or local law.** We encourage users to report content that they are concerned about. Once reported, we assess these reports and action the content in line with our policies.

We constantly update our tools and reporting channels so that users can more easily manage their experience and notify Meta if a piece of content or an account is concerning.

Below provides a reporting flow for impersonation, however, users can also use our in-app reporting to notify us of other harmful content including but not limited to misinformation, a fake account, bullying & harassment.

Impersonation in-app reporting in Facebook Messenger



During the 2022 Australian federal election campaign, we also put in place a number of additional arrangements with the AEC to ensure we had clear channels of communication during the campaign, including onboarding the AEC to a reporting system with 24 / 7 capability, and dedicated response teams.

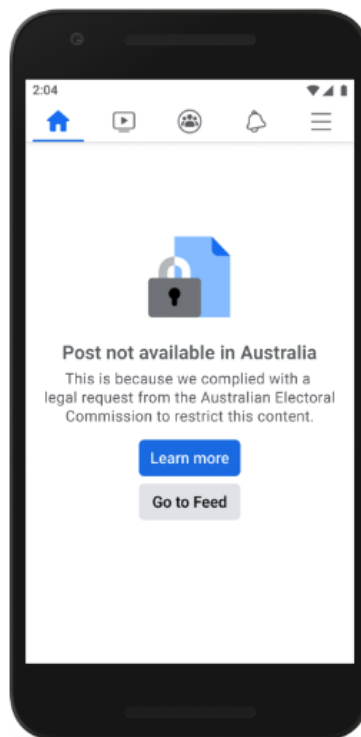
This is in addition to the referral pathways we already have in place with a number of agencies and government departments including the Department of Home Affairs, Department of Communications, the Australian Communications and Media Authority, and the Office of the eSafety Commissioner.

Through this channel, we review referred content first against our Community Standards. If we determine that the content goes against our policies, we remove it. If content does not go against our policies, in line with our commitments as a member of the Global Network Initiative and our Corporate Human Rights Policy, we conduct a careful legal review to confirm whether the report is valid, as well as human rights due diligence.³⁶

We provide notice to people when we restrict something they posted based on a report that the content goes against local law, and we also tell people when they try to view something that is restricted in their country on the basis of a takedown request.

³⁶ Meta, 'How we assess reports of content violating local law', *Transparency Centre*, <https://transparency.fb.com/data/content-restrictions/content-violating-local-law/>

Notice to users whose content is restricted due to a government takedown request



Hate speech

Hate speech is traditionally one of the most challenging types of content to proactively detect because it is so context-dependent. However, our investment in artificial intelligence is evident from the increasing percentage of hate speech content we have been detecting proactively.

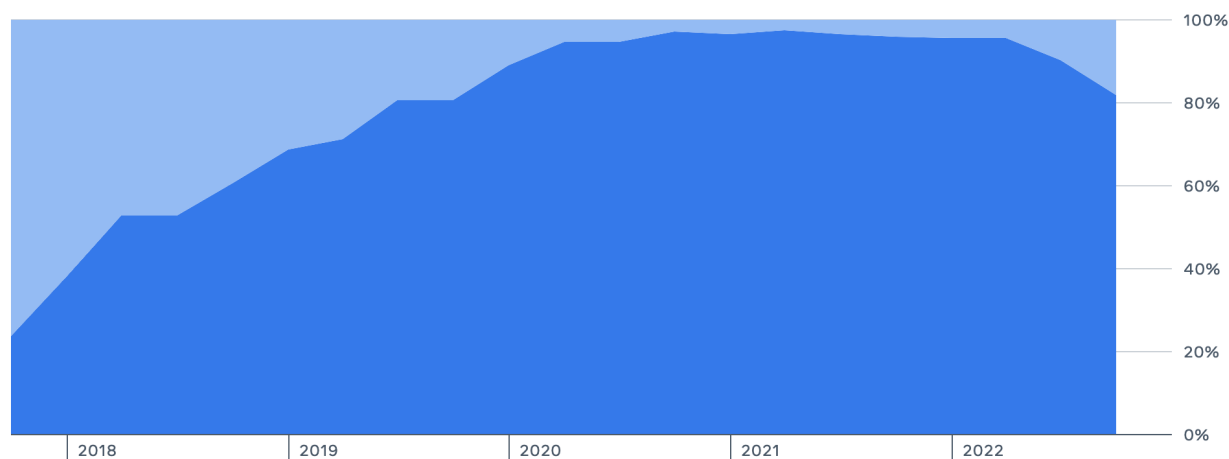
At the end of 2017, less than 25 per cent of hate speech content we removed was detected proactively. This figure has progressively increased over that time: by end 2018, 40 per cent was proactively detected; by end 2019, 80 per cent was proactively detected.

According to the latest Community Standards Enforcement Report, in the period October to December 2022, we took action against 11 million pieces of content for hate speech, of which 81.9 per cent was detected proactively.³⁷

³⁷ Meta, *Community Standards Enforcement Report Q2 2022 - Hate speech*
<https://transparency.fb.com/data/community-standards-enforcement/>

We also measure how prevalent violating content is on our services, that is, the number of views of violating content, divided by the estimated number of total content views on Facebook or Instagram.³⁸ At the end of 2020 0.7 to 0.8 per cent of views on Facebook contained hate speech. This means, for every 10,000 views of content on Facebook, 7 or 8 contained hate speech.³⁹ Now, as reported in December 2022, less than 0.02 per cent of views on Facebook contained hate speech.

Percentage of hate speech content proactively removed on Facebook, 2018 - 2022



During the 2022 Australia election campaign (between April 1 and June 30, 2022), Meta took action on over **91,000 pieces of content on Facebook** and over **40,000 pieces of content on Instagram** in Australia for violating our Community Standards on hate speech.

Violence and incitement

Between October to December 2022, globally, we took action on 13.1 million pieces of violence and incitement content on Facebook, and 87.3 percent of this content was proactively detected.⁴⁰

³⁸ Meta, Prevalence, *Transparency Centre*, <https://transparency.fb.com/en-gb/policies/improving/prevalence-metric/>

³⁹ A Kantor, Measuring our progress combatting hate speech, *Meta Newsroom*, 19 November 2020, <https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/>

⁴⁰ Meta, *Community Standards Enforcement Report Q2 2022 - Violence and incitement*, <https://transparency.fb.com/data/community-standards-enforcement/violence-incitement/facebook/>

Prevalence of violence and incitement content is 0.02 per cent, which means that for every 10,000 views of content on Facebook, 2 contained violence and incitement content.

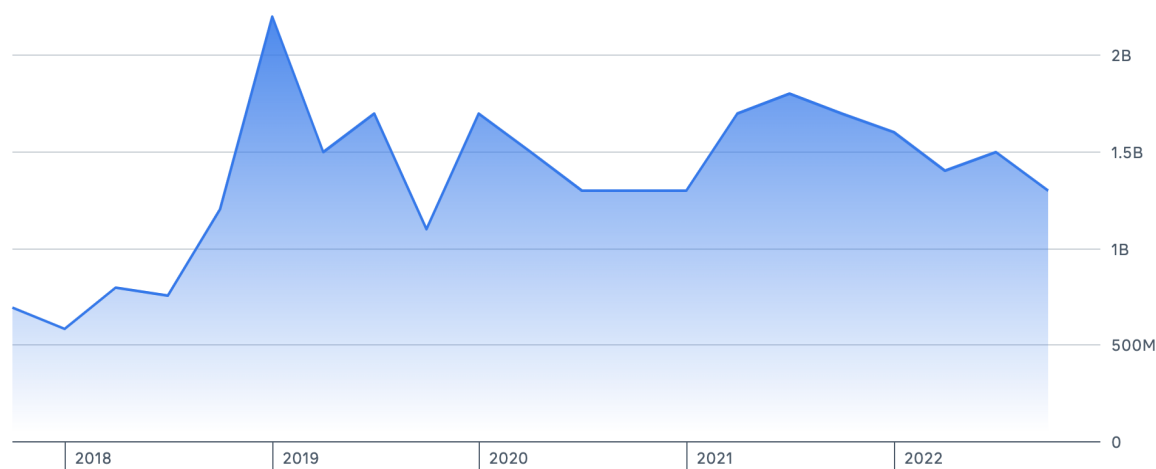
During the 2022 Australia election campaign (between April 1 and June 30, 2022), Meta took action on over **200,000 pieces of content on Facebook** and over **46,000 of content on Instagram** in Australia for violating our Community Standards on violence and incitement.

Fake accounts

Our ability to detect and remove fake accounts has been improving over the years, and there has been a general decline in the volume of fake accounts found on the platform since 2019.

Our most recent data, found that in October to December 2022 we globally removed 1.3 billion fake accounts on Facebook, and 99.3 percent of these were removed proactively.⁴¹

Number of fake accounts we've taken action on (2018-2022)



⁴¹ Meta, *Community Standards Enforcement Report Q2 2022 - Fake accounts*, <https://transparency.fb.com/data/community-standards-enforcement/fake-accounts/facebook/>

Misinformation

Meta continues to remove misinformation that violates our Community Standards,⁴² including misinformation that is likely to directly contribute to the risk of imminent physical harm.

In July 2022, we reported that we have removed more than 25 million pieces of content from Facebook and Instagram globally for violating our policies on COVID-19 related misinformation.⁴³

During the 2022 Australia election campaign (between April 1 and June 30, 2022), Meta took action on over **25,000 pieces of content across Facebook & Instagram** for violating our Harmful Health Misinformation policies.

We displayed warnings on over **3 million distinct pieces of content** on Facebook (including reshares) based on articles written by our third party fact checking partners.

Coordinated inauthentic behaviour

We report regularly on our approach to CIB to provide the community, civil society and governments with greater confidence in our efforts to combat these operations. These are reported through:

- **Our Community Standard Enforcement Report.** Each quarter, we report on metrics for preventing and taking action on content that goes against our Community Standards.⁴⁴
- **Monthly Adversarial Threat reports.** Each month we publish a list of CIB networks that we have taken down.⁴⁵

⁴² Meta, *Community Standards*, <https://transparency.fb.com/en-gb/policies/community-standards/>

⁴³ N Clegg, 'Meta asks Oversight Board to advise on COVID-19 misinformation policies', *Meta Newsroom*, 26 July 2022, <https://about.fb.com/news/2022/07/oversight-board-advise-covid-19-misinformation-measures/>

⁴⁴ Meta, *Community Standards Enforcement Report*, <https://about.fb.com/news/tag/coordinated-inauthentic-behavior/>

⁴⁵ Meta, *Community Standards - Coordinated Inauthentic Behaviour*, <https://about.fb.com/news/tag/coordinated-inauthentic-behavior/>

The United States was the most targeted country by global CIB operations we've disrupted over the years, followed by Ukraine and the United Kingdom.

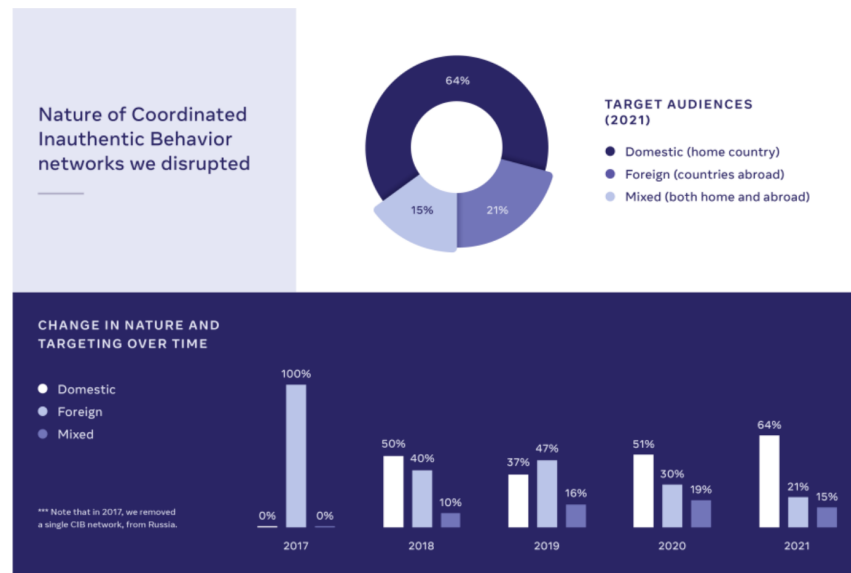
GLOBAL CIB DISRUPTIONS, 2017-2022

(by country of origin)

A world map visualization showing the geographic distribution of CIB (Creative Intellectual Property) disruptions from 2017 to 2022. The map uses a light blue background with darker blue dots representing the locations of disruptions. The dots are concentrated in North America, Europe, and Asia, with a significant increase in density over time. A timeline at the bottom of the map shows the years 2017, 2018, 2019, 2020, 2021, and 2022, with a corresponding bar chart indicating the volume of disruptions for each year.

⁴⁶ B Nimmo, D Agranovich, 'Recapping our 2022 coordinated inauthentic behaviour enforcements', *Meta Newsroom*, 15 December 2022, <https://about.fb.com/news/2022/12/metas-2022-coordinated-inauthentic-behavior-enforcements/> and Meta, 'December 2021 Coordinated Inauthentic Behaviour Report', December 2021, <https://about.fb.com/wp-content/uploads/2022/01/December-2021-Coordinated-Inauthentic-Behavior-Report-2.pdf>

Target of Coordinated Inauthentic Behaviour Disruptions, 2017 - 2022⁴⁷



We also recently reported that we had identified more than 400 malicious android and iOS apps that were designed to steal Facebook login information and compromise people's accounts.⁴⁸ These apps were listed on the Google Play Store and Apple's App Store and disguised as photo editors, games, VPN services, business apps and other utilities to trick people into downloading them.

We know that CIB threats are rarely confined to one platform. We share our findings and threat indicators with industry peers so they too can detect and stop threat activity, and we can build our collective response to CIB.

During the Australian election, we did not see any evidence of CIB targeting Australia.

⁴⁷ We define targets as:

- Domestic: IO that targets public debate in the same country from which it operates.
- Foreign: IO that targets the public debate in a different country from which it operates.
- Mixed: We also see IO campaigns and threat actors that run campaigns that target both domestic and foreign audiences

⁴⁸ D Agranovich, 'Protecting people from malicious account compromise apps', *Meta Newsroom*, 7 October 2022, <https://about.fb.com/news/2022/10/protecting-people-from-malicious-account-compromise-apps/>

Partnerships

Safeguarding the integrity of elections requires cross-sector collaboration. Meta continues to partner with industry, government, academics and civil society organisations to ensure the measures we take to enhance the safety and security and inform our users about the election process have the most effective impact.

Research partnerships

We continue to invest in research to better understand the nuances of misinformation, disinformation and coordinated inauthentic behaviour which may influence an election campaign. Some recent highlights include:

- Commissioning independent research by respected Australian academic Dr Andrea Carson to map government approaches to combatting misinformation around the world, focussing on the Asia-Pacific region.⁴⁹ The report ‘Tackling Fake News’ was launched in January 2021 and has been positively received by policymakers and experts across the world as they consider new approaches to combating misinformation.
- Investing in academic research on misinformation and polarisation. In 2022 Meta released a request for proposals (RFP) for research exploring integrity issues related to social communication technologies. The RFP attracted 503 proposals from 349 universities and institutions around the world.

We provided a total of \$1,000,000 USD in funding for 11 proposals, including research from the Queensland University of Technology by Michelle Riedlinger and Silvia Montaña-Niño, and Marina Joubert (Stellenbosch University), Víctor García-Perdomo (Universidad de La Sabana) on ‘Countering misinformation in the Southern Hemisphere: A comparative study’.⁵⁰

⁴⁹ A Carson, ‘*Fighting Fake News: A Study of Online Misinformation Regulation in the Asia-Pacific*’, https://www.latrobe.edu.au/_data/assets/pdf_file/0019/1203553/carson-fake-news.pdf

⁵⁰ Meta Research, ‘Announcing the winners of the 2022 foundational integrity research request for proposals’, *Meta Research*, 28 March 2023, <https://research.facebook.com/blog/2023/2/announcing-the-winners-of-the-2022-foundational-integrity-research-request-for-proposals/>

In 2021, we supported two Australian universities:

- ‘Testing fact and logic-based responses to polarising climate misinformation’ (John Cook and Sojung Kim, Monash University); and
 - ‘How fact checkers compare: News trust and COVID-19 information quality’ (Andrea Carson, James Meese, Justin B. Phillips, Leah Ruppanner, La Trobe University).⁵¹
- Meta supported an analytical paper by First Draft (now RMIT FactLab) on disinformation and misinformation amongst diaspora groups with a focus on Chinese language.⁵² The paper aims to inform policymakers on how to reduce misinformation within Chinese diaspora communities.
 - We are a major sponsor of the Australian Strategic Policy Institute (ASPI). We also funded Dr Jake Wallis from ASPI to undertake a review of disinformation-for-hire, specifically targeting Australia and the Asia-Pacific region. This research was launched in August 2021.⁵³
 - Sponsoring the Australian Media Literacy Alliance’s Media Literacy Summit. Meta has sponsored the Australian Media Literacy Alliance to host the inaugural Australian Media Literacy Summit, which took place in March 2023.

The Summit focussed on building a network of media literacy advocates and leaders who will work toward developing media literacy in schools and communities across Australia.

- Sponsoring the Australian Strategic Policy Institute’s ‘Sydney Dialogue’ which took place in April 2023. This was a high level two day event to discuss significant technology issues, including misinformation and disinformation.

⁵¹ Meta Research, ‘Announcing the 2021 recipients of research awards in misinformation and polarisation’, *Meta Research*, 14 September 2021, <https://research.facebook.com/blog/2021/09/announcing-the-2021-recipients-of-research-awards-in-misinformation-and-polarization/>

⁵² E Chan, S Zhang, ‘Disinformation, stigma and chinese diaspora: policy guidance for Australia’, *First Draft website*, 31 August 2021, <https://firstdraftnews.org/long-form-article/disinformation-stigma-and-chinese-diaspora-policy-guidance-for-australia/>

⁵³ Dr J Wallis, ‘Influence for hire: the Asia-Pacific’s online shadow economy’, *Australian Strategic Policy Institute*, 10 August 2021, <https://www.aspi.org.au/report/influence-hire>

Public awareness and media literacy campaigns

We also partner with experts to develop media literacy campaigns and raise awareness about misinformation.

- In September 2021, we launched a “Don’t Be a Misinfluencer” campaign with First Draft (now RMIT FactLab) for public figures and creators. The campaign aimed to prevent the amplification of misinformation by creators and includes a toolkit with information on how to identify and combat misinformation.⁵⁴

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



- In October 2021, we launched a media literacy ‘Check the Facts’ initiative for Australians with the Australian Associated Press to teach Australians how to recognise and avoid the spread of misinformation.⁵⁵

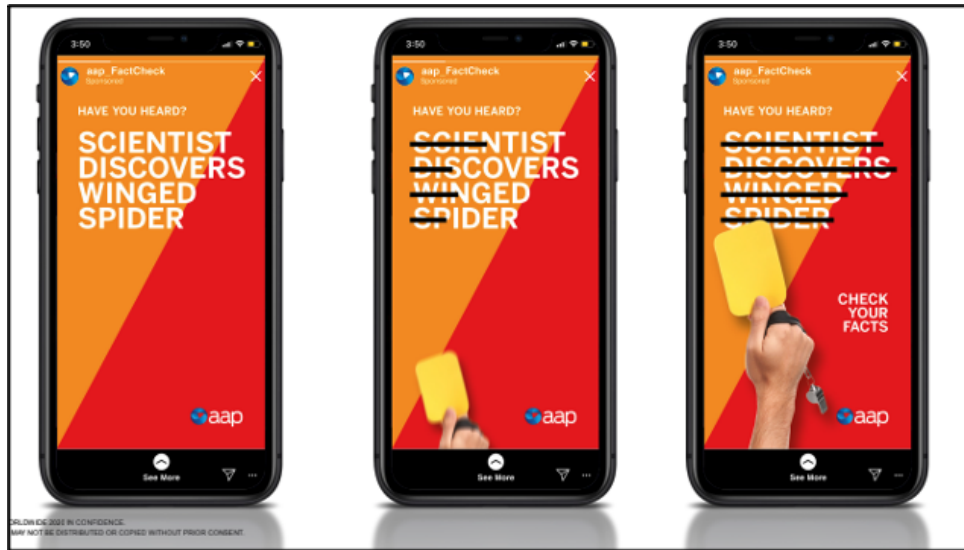
We translated this campaign into Chinese, Arabic and Vietnamese in the lead-up to the Australia election campaign, to spread awareness about misinformation in culturally diverse communities.

We ran this campaign twice - once in the second half of 2021, and then again before the election in May 2022. In total, the campaign and its assets were seen by 23.9 million people across Australia.

⁵⁴ First Draft, ‘Protect your voice: a toolkit for Australian influencers and celebrities’, *First Draft website*, <https://firstdraftnews.org/tackling/protect-your-voice-a-toolkit-for-australian-influencers-and-celebrities/>

⁵⁵ Australian Associated Press, ‘Australians urged to Check the Facts’, *AAP website*, 25 October 2021, <https://www.aap.com.au/news/australians-urged-to-check-the-facts/>

Australian Associated Press 'Check the Facts' Campaign



Working with government and security agencies

We share our findings and threat indicators with industry peers so they too can detect and stop threat activity, and we can build our collective response to CIB.

In particular, in the lead up to the Australia election, we worked with the Australian Government's EIAT and developed responses to a series of possible scenarios that may arise during the course of an election campaign.

We also engaged with Australian security agencies in the year leading up to the Australian election campaign.

Transparency and accountability

Finally, we recognise the importance of giving users transparency and control around what they see on our services, and our approach to content governance, and we provide a number of resources to increase transparency in this area.

While this submission focuses primarily on the actions we take in relation to safeguard the integrity of elections, it is worth noting the considerable efforts Meta puts in place to enhance transparency and accountability for the decisions we take generally.

Below provides an overview of the transparency tools we provide for organic content and advertising for political purposes, as well as the Transparency Reports and Oversight Mechanisms we outline in our Transparency Centre⁵⁶ which gives our community visibility over how we enforce our policies.

Political advertising and potential foreign interference

Meta provides industry-leading transparency of political advertising on our platform. We believe these steps help to promote an informed electorate.

These steps also help to increase transparency and authenticity around foreign influence - noting that foreign influence differs from foreign interference. When foreign entities aim to affect the political debate within Australia but do so openly and transparently, this is better described as foreign influence. Foreign actors can have a legitimate role in participating in Australian political debates. However, we recognise that there is value in providing transparency about the physical location of entities engaged in those debates to give confidence about the role those entities play in the political process.

Over the years we have updated our policies and tools to increase transparency in a number of ways:

- **Political, electoral and social issue ad transparency.** Meta now requires all advertisers of social issue, electoral and political (SIEP) ads in Australia to complete a number of steps to confirm their authenticity and enhance transparency.⁵⁷

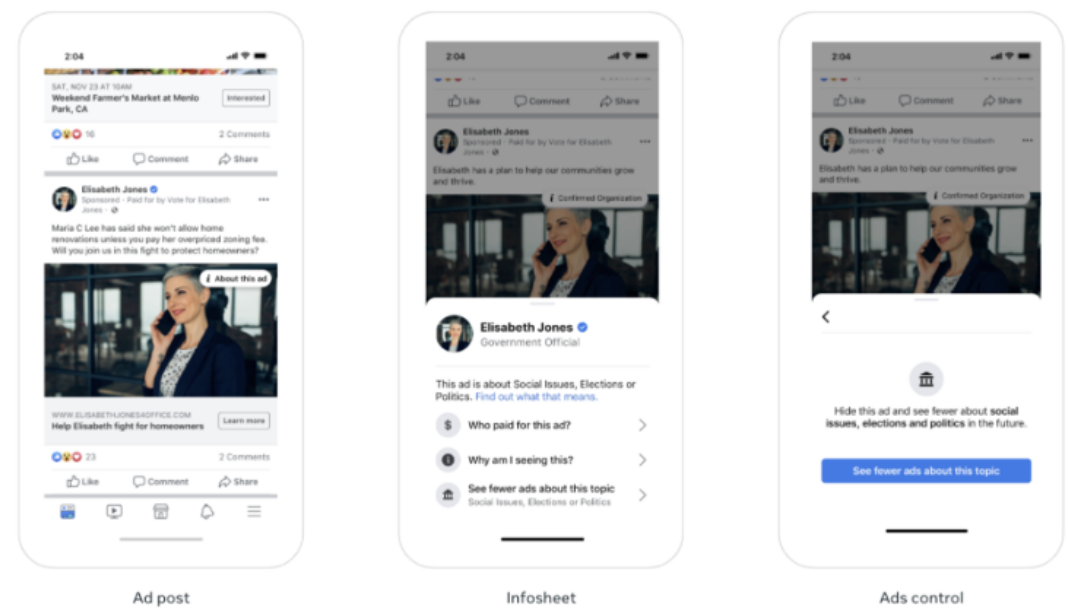
⁵⁶ Meta, *Transparency Centre*, <https://transparency.fb.com/data/>

⁵⁷ J Machin, 'Expanding transparency around social issue ads in Australia', *Meta Australia Blog*, 18 June 2021, <https://medium.com/meta-australia-policy-blog/expanding-transparency-around-social-issue-ads-in-australia-c71f8e26d407>

Advertisers of SIEP ads must provide identification and be authorised by Meta prior to running the ad. Advertisers must also include a “Paid for by” disclaimer on their ad, and have their ads stored in Meta’s publicly available Ad Library for seven years, even if the page that posted them is no longer operational.⁵⁸

During the Australian election campaign we rejected around **17,000** ads for not complying with our political and social issue ads enforcement policies.

Political ad transparency disclaimer



- **Ad Library and Ad Library report.** Meta has also established an Ad Library to provide additional transparency around social issues, political and election ads on our services.

The Ad Library is an industry-leading transparency initiative, which provides information to anyone about the ads on Facebook and Instagram, and who is behind them.

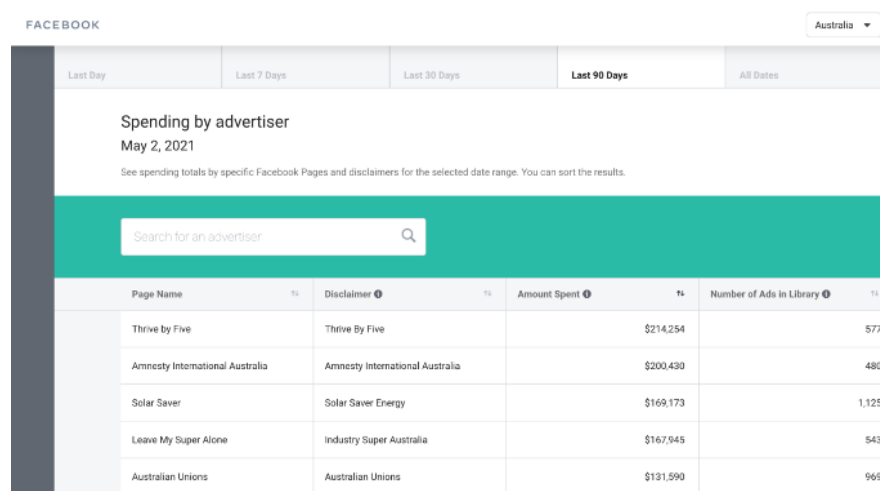
⁵⁸ Meta, *Ad Library*,
https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=AU

- We have progressively added new features to the Ad Library. We now include all active ads along with more Page information such as creation date, name change, Page mergers and the primary country of people who manage Pages with large audiences. We also provide a summary of targeting information at the Advertiser level that includes the total number of ads run, targeting types (location, demographic, interests), and whether the Page used customer or lookalike audiences.⁵⁹

To enable easier comparison and analysis between advertisers we also provide the Ad Library Report which allows for aggregated analysis of political ads on Facebook and Instagram.

We also now allow an Ad Library API which encourages greater scrutiny of advertisers on Facebook and Instagram.

Meta Ad Library



FACEBOOK Australia

Last Day Last 7 Days Last 30 Days **Last 90 Days** All Dates

Spending by advertiser
May 2, 2021

See spending totals by specific Facebook Pages and disclaimers for the selected date range. You can sort the results.

Search for an advertiser

Page Name	Disclaimer	Amount Spent	Number of Ads in Library
Thrive by Five	Thrive By Five	\$214,254	577
Amnesty International Australia	Amnesty International Australia	\$200,490	480
Solar Saver	Solar Saver Energy	\$169,173	1,125
Leave My Super Alone	Industry Super Australia	\$167,945	543
Australian Unions	Australian Unions	\$131,590	969

- **Instituting a greater level of transparency to Facebook Pages.**⁶⁰ Under the Transparency tab of Pages, users are able to see the location of Page admins, the history of a Page, and ads currently running.

⁵⁹ J King, 'Bringing more transparency to social issue, electoral and political ads, *Meta Business*, 23 May 2022, <https://www.facebook.com/business/news/transparency-social-issue-electoral-political-ads>

⁶⁰ Meta, 'A New Level of Transparency for Ads and Pages', *Meta Newsroom*, 28 June 2018, <https://newsroom.fb.com/news/2018/06/transparency-for-ads-and-pages/>

- **Applying labels to content from media outlets that are partly or fully under the control of their government.** We provide greater transparency into these publishers because they combine the influence of a media organisation with the strategic backing of a state, and we believe people should know if the news they read is coming from a publication that may be under the influence of a government.⁶¹

Transparency reports

It is important to us that we maintain transparency with the community and policymakers around how we make content decisions, and the nature and extent of the government requests we receive for user data. We detail this information in a number of reports, including:

Community Standards Enforcement Report

Each quarter we release a Community Standards Enforcement Report (CSER). Our latest CSER was published in December 2022 and covers the period between October and December 2022.⁶²

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

1. **Content removed** for going against our standards. This metric shows the scale of our enforcement activity.
2. **Content removed proactively** before users reported them to us. We use this metric as an indicator of how effectively we detect violations.
3. **Prevalence.** Prevalence metrics allow us to track, both internally and externally, how much violating content people are seeing on our apps. Prevalence, in turn, helps us determine the right approaches to driving that metric down, whether it's through updating our policies, products or tools for our community.
4. **Appeals.** We report the number of pieces of content that people appeal after we take action on it for going against our policies. Reporting on this metric holds us to

⁶¹ N Gleicher, 'Labeling State-Controlled Media', *Meta Newsroom*, 4 June 2020, <https://about.fb.com/news/2020/06/labeling-state-controlled-media/>

⁶² Meta, *Community Standards Enforcement Report* - <https://transparency.fb.com/data/community-standards-enforcement/>

account for our content decisions, and ensures we can continue to improve our enforcement.

5. **Restored content.** For policy violations, we measure the number of pieces of content that we restored after we originally took action on them.

To ensure our CSER remains an accurate and meaningful measure of Meta's content moderation, we appointed the Data Transparency Advisory Group (DTAG).⁶³ The DTAG is an independent body made up of international experts in measurement, statistics, criminology and governance who provide independent, public assessments of the CSER metrics, and the enforcement processes and measurement methodologies we use.

In 2019, the DTAG completed a formal review of the CSER.⁶⁴ They found that Meta's approach to content moderation processes are appropriate given the scale at which we operate and the amount of content people post. They also found that the accuracy of our content review system was well designed, and the metrics we use to measure success (prevalence, actioned content and proactive rate), are in line with best practice.

DTAG also laid out 15 recommendations which Meta has continued to implement. These include additional metrics we should report on, further break-downs of the metrics we already provide, and making it easier for people who use Meta's services to stay updated on the changes we make to our policies.⁶⁵

In 2020, we committed to undergoing an independent audit, conducted by EY, to validate that our metrics published in the Community Standards Enforcement Report are measured and reported correctly. As part of the assessment, Meta provided EY with full access to the necessary data, documentation and evidence requests. We also gave access to dozens of employees across data science, data engineering, software engineers, product and program managers and Internal Audit teams working on the Community Standards Enforcement Report.

⁶³ Meta, 'An independent report on how we measure content and moderation', *Meta Newsroom*, 23 March 2019, <https://about.fb.com/news/2019/05/dtag-report/>

⁶⁴ Meta, 'An independent report on how we measure content and moderation', *Meta Newsroom*, 23 March 2019, <https://about.fb.com/news/2019/05/dtag-report/>

⁶⁵ The full Report of the Facebook Data Transparency Advisory Group can be found here https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf

In May 2022, EY published its assessment and found the calculation of the metrics in our 2021 fourth quarter Community Standards Enforcement Report were fairly stated, and our internal controls are suitably designed and operating effectively.⁶⁶

Other transparency reports

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

- **Content Distribution Guidelines.**⁶⁷ As mentioned above, these Guidelines outline what content receives reduced distribution on Feed because it's problematic or low quality.
- **Recommendation Guidelines.** Meta publishes a set of Recommendation Guidelines, both on Facebook⁶⁸ and Instagram,⁶⁹ which outline the types of content that may not be eligible for recommendations.⁷⁰
- **Government requests for user data.**⁷¹ Meta produces this report to provide information on the nature and extent of government user data requests.
- **Content restrictions.**⁷² This report details instances where we limited access to content based on local law.
- **Intellectual property report.**⁷³ This report details how many reports of intellectual property violations we received and how much content we took down as a result.
- **Adversarial Threat Report.**⁷⁴ Each quarter we publish a list of coordinated inauthentic behaviour networks that we have taken down.

⁶⁶ Meta, *Community Standards Enforcement Report*, <https://about.fb.com/news/2022/05/community-standards-enforcement-report-assessment-results/> The results of the EY assessment can be found at <https://about.fb.com/wp-content/uploads/2022/05/EY-CSER-Independent-Assessment-Q4-2021.pdf>

⁶⁷ 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/>

⁶⁸ Meta, 'What are recommendations on Facebook?', *Help Centre*, <https://www.facebook.com/help/1257205004624246>

⁶⁹ Instagram, 'What are recommendations on Instagram?', *Help Centre*, <https://help.instagram.com/313829416281232>

⁷⁰ G Rosen, 'Recommendation guidelines', *Meta Newsroom*, 31 August 2020, <https://about.fb.com/news/2020/08/recommendation-guidelines/>

⁷¹ Meta, *Government Requests for User Data*, <https://transparency.fb.com/data/government-data-requests/>

⁷² Meta, *Content Restrictions*, <https://transparency.fb.com/data/content-restrictions/>

⁷³ Meta, *Intellectual Property*, <https://transparency.fb.com/data/intellectual-property/>

⁷⁴ N Gleicher, 'Meta's Adversarial Threat Report', *Meta Newsroom*, 1 December 2021, <https://about.fb.com/news/2021/12/metas-adversarial-threat-report/>

- **Widely viewed content report.**⁷⁵ The WVCR is released quarterly and aims to provide more transparency and context about what people are seeing on Facebook by sharing the most-viewed domains, links, Pages and posts for a given quarter in Feed in the United States.
- **Facebook Open Research and Transparency (FORT) initiative.** Finally, in November 2021 we announced a new Researcher API as part of the Facebook Open Research and Transparency (FORT) initiative, which allows qualified academics to pull reports and conduct longitudinal research across all public Facebook Pages, Groups and Events.⁷⁶ This API will help academics understand how public discussions on Facebook influence the social issues of the day.

In July 2022, we expanded this initiative to provide detailed targeting information for social issue, electoral or political ads to vetted academic researchers through the Facebook Open Research and Transparency (FORT) environment.⁷⁷

Independent oversight

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

The Oversight Board was borne out of the recognition that critical decisions about content should not be left to companies alone. Content decisions can have significant consequences for free expression and companies like Meta - notwithstanding our significant investments in detection, enforcement and careful policy development - will not always get it right.

The Oversight Board comprises 23 experts in human rights and technology - including the Australian academic Professor Nic Suzor from Queensland University of Technology. The Board is entirely independent and hears appeals on Meta's decisions relating to

⁷⁵ A Stepanov, 'Introducing the Widely Viewed Content Report', 18 August 2021, *Meta Newsroom*, <https://about.fb.com/news/2021/08/widely-viewed-content-report/>

⁷⁶ G Rosen, 'Community standards enforcement report - third quarter 2021', *Meta Newsroom*, 9 November 2021, <https://about.fb.com/news/2021/11/community-standards-enforcement-report-q3-2021/>

⁷⁷ J King, 'Bringing more transparency to social issue, electoral and political ads', *Meta Business*, 23 May 2022, <https://www.facebook.com/business/news/transparency-social-issue-electoral-political-ads>

content on Facebook and Instagram. We have agreed that the Board's decisions will be binding, and the Board is also able to make recommendations about Meta's policies.⁷⁸

The Oversight Board began issuing decisions in January 2021.⁷⁹ Since publishing its first decisions in January 2021, the Oversight Board has made more than 100 recommendations to Meta for future improvements.⁸⁰

The Oversight Board has also recently begun publishing Transparency Reports which provide new details on the Oversight Board's cases, decisions and recommendations.

As of December 2022, the Oversight Board had issued 36 case decisions and 176 recommendations to Meta for future improvements. Meta has reported its progress against implementing 140 of these recommendations. Meta has implemented 24 (17%) of the Board's recommendations fully, as demonstrated through published information. Eleven, (8%) have been partially implemented, and Meta has reported progress towards implementing 53 (38%). Meta has reported implementation against 28 (20%) recommendations, or said it already does what the Board recommends.⁸¹

Most recently, Meta asked the Oversight Board for advice on whether current measures to address dangerous COVID-19 misinformation on the platform should remain in place.⁸²

Misinformation related to COVID-19 has presented unique risks to public health and safety over the last two years and more. To keep our users safe while still allowing them to discuss and express themselves on this important topic, we broadened our harmful misinformation policy in the early days of the outbreak in January 2020. The change meant that, for the first time, the policy would provide for removal of entire categories of false claims on a worldwide scale.

⁷⁸ B Harris, 'Establishing structure and governance for an independent oversight board', *Meta Newsroom*, 17 September 2019, <https://about.fb.com/news/2019/09/oversight-board-structure/>

⁷⁹ N Clegg, 'Welcome the oversight board', *Meta Newsroom*, 6 May 2020, <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/>

⁸⁰ Oversight Board, 'Oversight Board demands more transparency from Facebook', *Oversight Board*, October 2021, <https://oversightboard.com/news/215139350722703-oversight-board-demands-more-transparency-from-facebook/>

⁸¹ Oversight Board, 'Oversight Board announced plans to review more cases, and appoints a new board member', *Oversight Board*, February 2023, <https://www.oversightboard.com/news/943702317007222-oversight-board-announces-plans-to-review-more-cases-and-appoints-a-new-board-member/>

⁸² N Clegg, 'Meta asks Oversight Board to advise on COVID-19 misinformation policies', *Meta Newsroom*, 26 July 2022, <https://about.fb.com/news/2022/07/oversight-board-advise-covid-19-misinformation-measures/>

The pandemic has evolved. In many countries, where vaccination rates are relatively high, life is increasingly returning to normal. But this isn't the case everywhere and the course of the pandemic will continue to vary significantly around the globe — especially in countries with low vaccination rates and less developed healthcare systems.

It is important that any policy Meta implements be appropriate for the full range of circumstances countries find themselves in. Meta therefore sought advice from the Oversight Board about our measures to address COVID-19 misinformation, including whether those introduced in the early days of an extraordinary global crisis remain the right approach, or whether we should address this misinformation through other means, like labelling or demoting it either directly or through our third-party fact-checking program.

In April 2023, the Board issued 18 recommendations, including for Meta to continue to remove harmful misinformation about COVID-19 for as long as the World Health Organization declares COVID-19 a global public health emergency. The Oversight Board also recommended Meta review the claims it removes under its Misinformation and Harm Policy in consultation with expert stakeholders.⁸³

Meta is preparing our response to the Oversight Board's recommendations. We will continue to enforce on our Misinformation and Harm policy while we prepare our response.

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

We welcome the opportunity to discuss Meta's efforts to safeguard the integrity of the upcoming referendum further, and look forward to collaborating with the Australian Government and broader industry on how to protect the integrity of political debate on our platform in the lead up to the referendum.

⁸³ Oversight Board, 'Oversight Board publishes policy advisory opinion on the removal of COVID-19 misinformation', *Oversight Board*, April 2023, <https://oversightboard.com/news/739141534555182-oversight-board-publishes-policy-advisory-opinion-on-the-removal-of-covid-19-misinformation/>