



Australian Code of Practice on Disinformation and Misinformation
Adobe, Inc.
Annual Transparency Report
January 2023 – December 2023

Summary

Adobe is pleased to continue its participation in the Australian Code of Practice on Disinformation and Misinformation.

Adobe is a global leader in digital marketing and digital media solutions. Since the company's foundation in December 1982, we have pushed the boundaries of creativity with products and services that allow our customers to create, deploy, and enhance digital content. Our purpose is to serve the creator and respect the consumer, and our heritage is built on providing trustworthy and innovative solutions to our customers.

With the increasing volume and velocity of digital content creation, including synthetic media, it is critical to ensure transparency, understanding, and trust in what we are consuming online while empowering consumers. Adobe feels a responsibility to support the creative community, and society at large, and is committed to finding solutions that help address the issues of manipulated media and tackle misinformation and disinformation.

As such, content provenance is a major focus for Adobe and the work we lead on the Content Authenticity Initiative (CAI). We are focused on cross-industry participation, with an open, extensible approach for providing media transparency to allow for better evaluation of content.

The CAI advocates for a set of open standards that can be used to create and reveal provenance for images, documents, time-based media (video, audio) and streaming content. Provenance, sometimes referred to as attribution, empowers content creators, editors, and publishers, regardless of their geographic location or degree of access to technology, to voluntarily disclose information about who created or changed an asset, what was changed and how it was changed.

In February 2021, Adobe, Arm, BBC, Intel, Microsoft, and Truepic launched a formal coalition for standards development: The Coalition for Content Provenance and Authenticity (C2PA). The C2PA is a mutually governed consortium created to accelerate the pursuit of pragmatic, adoptable standards for digital provenance, serving creators, editors, publishers, media platforms, and consumers.

In January 2022, the C2PA publicly released the open technical specification for digital provenance, which provides platforms with a blueprint to define what information is associated with each type of asset (e.g. images, videos, audio, or documents), how that information is presented and stored, and how evidence of tampering can be identified.

In March 2023, Adobe's new generative AI model, [Firefly](#), was announced, and along with it our commitment to leveraging CAI's Content Credentials to bring transparency to generative AI outputs. Every asset produced with Firefly has embedded a Content Credential indicating the model used and its version. This is significant — it not only builds on our mission to ensure tools like Firefly are used responsibly, but also gives viewers of this content important context to understand what they're seeing or hearing, enabling them to make trust decisions when necessary.

2023 was the most vibrant year yet for the CAI, C2PA and Content Credentials, and we can't wait to build on the momentum alongside our members and industry partners to ensure an even greater and lasting impact in 2024 and beyond. As of April 2024, we are at over 3,000 members in the Content Authenticity Initiative.

When we come together across technology, government, and civil society, listening to creators and information consumers, we bolster a basic right for everyone to understand how the content they consume was made.

<p>Objective 1 Safeguards against Disinformation and Misinformation:</p> <p>Outcome 1a: Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures.</p> <p>Specifically measures implemented under 5.9:</p> <ul style="list-style-type: none">H. the provision or use of technologies which assist digital platforms or their users to check authenticity or accuracy or to identify the provenance or source of digital content;I. exposing metadata to users about the source of content;
<p>Objective 3 Work to ensure the integrity and security of services and products delivered by digital platforms:</p> <p>Outcome 3: The risk that Inauthentic User Behaviours undermine the integrity and security of services and products is reduced.</p> <p>Please see Objective 1a.</p>
<p>Objective 4 Empower consumers to make better informed choices of digital content:</p> <p>Outcome 4: Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation.</p> <p>5.21. Signatories will implement measures to enable users to make informed choices about Digital Content and to access alternative sources of information.</p> <p>Specifically, measures developed and implemented in accordance with the commitment:</p> <ul style="list-style-type: none">C. the provision or use of technologies which signal the credibility of news sources, or which assist digital platforms or their users to check the authenticity or accuracy of online news content, or to identify its provenance or source;D. the promotion of digital literacy interventions, informed by evidence or expert analysis <p>Please also see Objective 1a.</p>

Objective 6 Strengthen public understanding of Disinformation and Misinformation through support of strategic research:

Outcome 6: Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation.

Objective 7 Publicise the measures we take to combat Disinformation:

Outcome 7: The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.

Following the Code's guidance on signatories nominating to report on specified provisions in the Code, we have again this year reported on measures that are "proportionate and relevant" to our business. Adobe has considered the Code's guiding principles and the context in which our products and services might contribute to the harms arising from the spread of disinformation and misinformation on online platforms.

Adobe produces content creation and editing tools to help individuals and enterprises accelerate their productivity as they create, publish, and promote their creative work. While some of our products – most notably Behance – allow users to share, showcase, and promote their content online, Adobe is not a social media company. None of our products facilitate global conversations about current events or allow users to share and disseminate news content to global audiences. We believe digital creative works – the primary purpose of Adobe's suite of products – and any associated harms that stem from them are not the focus of the Code.

While Adobe's products and services fall outside the scope of the Code, we share the expressed concerns about harms that may result from malicious actors using our tools to produce inaccurate digital content. Therefore, we have opted into the provisions of the Code that focus on technologies we are developing to help users authenticate online media. To help mitigate the negative impact of misinformation and disinformation, Adobe is focused on providing tools to digital platforms that can help their users determine the sources and authenticity of online content.

Reporting against Commitments

Outcome 1a: Reducing harm by adopting scalable measures

At Adobe we have focused on four core areas to meet this Commitment.

1. Sharing open-source code with the community to help support adoption of provenance.
2. Implementing provenance technology in Adobe products so that our millions of customers can use this technology to show metadata about content created.
3. Providing a resource for audiences everywhere to check for provenance and attribution history.
4. Supporting CAI members with their integration of provenance using the open-source code and the C2PA standard.

Open Source

In June 2022, Adobe released a suite of open-source developer tools based on the C2PA specification, enabling more developers to integrate content provenance across web, desktop, and mobile projects

— for free. This is helping to get provenance tools into the hands of millions of creators and developers to create a safer, more transparent digital ecosystem, while providing users with information to be better informed about the content they see online.

Our team supports three options for open-source tools to implement C2PA standards beyond just Adobe apps. These tools include options from lightweight JavaScript to read Content Credentials on your site or app to completely customisable tools with the full SDK. Building on from our original release of the open-source tools, in early 2023 we updated them and now include both documentation and quicker file processing time to improve the developer experience. In addition to support for selected video and audio formats, and mobile development, as we continue to explore how we can expand content provenance to new mediums. Since the open-source tools were first published they have been downloaded over 8200 times.

Provenance Technology in Products – ex. Adobe Firefly

In 2023, the world saw that generative AI has the powerful ability to create new content in seconds using just a few keystrokes. It is transforming the way we work, create, and communicate. For example, generative AI allows you to generate convincing synthetic images of political leaders, celebrities, and other imagined scenes almost instantly. Across all types of generative AI content, it is becoming increasingly difficult to distinguish between fact and fiction. While we are optimistic about the technology and encourage its use, we also wanted to provide the tools to ensure the transparency needed to help create trust.

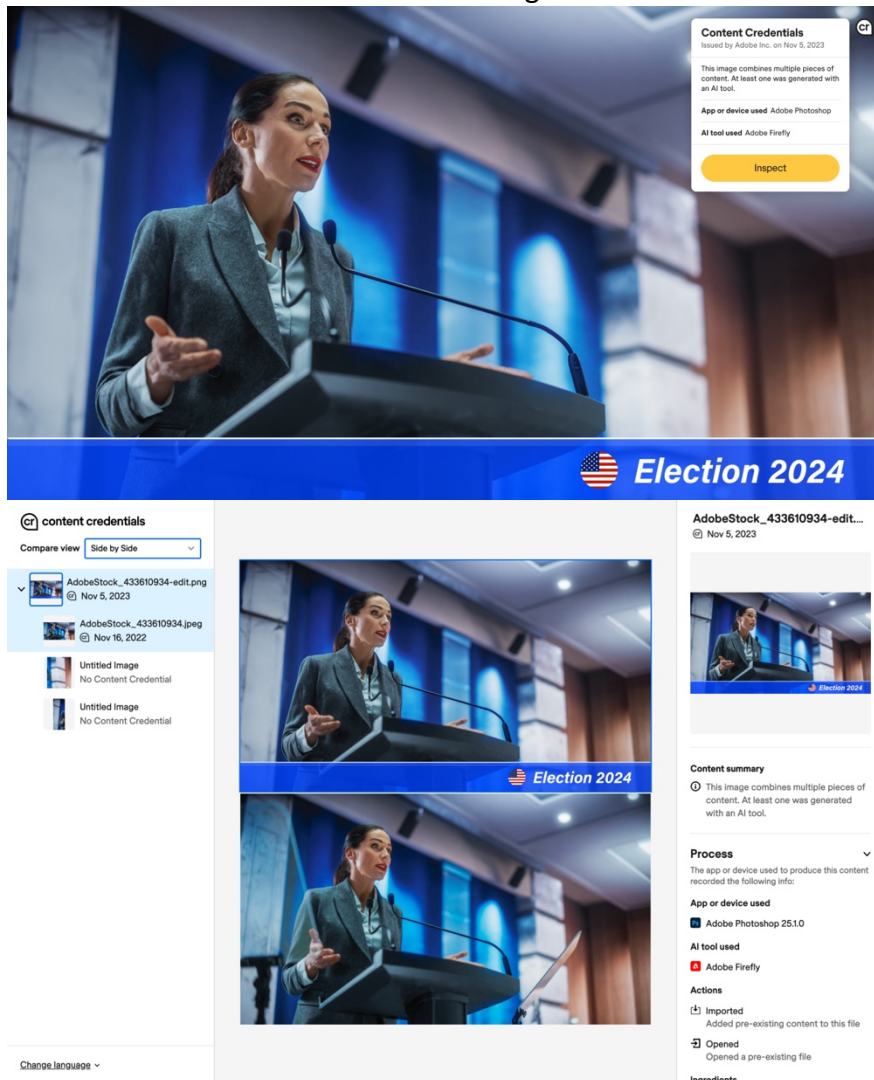
In March 2023, Adobe launched our own family of generative AI models, called Adobe Firefly. Since its launch, it has been used to generate over six billion assets. In keeping with our AI Ethics program, our leadership and product teams quickly made the decision that Content Credentials should be attached to content generated by Firefly, indicating that generative AI was used.

When a user creates and then downloads an asset in Firefly, a notification appears to the user to confirm that Content Credentials have been attached, letting people know that a generative AI tool has been used and which model. Content Credentials are then attached once the image is downloaded. The image can then be dropped into Adobe's inspection tool, "Verify," which will display this provenance information. Using open-source tools, publishers are able to add an interactive Content Credentials icon to appear within the digital content, allowing the viewer to see this provenance information with a simple mouse click

This decision was significant as it builds on our mission to ensure that AI tools like Firefly are used responsibly, giving viewers of digital content important context to help them understand what they are seeing and allowing them to make their own decisions about whether to trust the content.

In addition, we also transparently disclose to users of Firefly that Content Credentials will be attached. All users of Firefly agree to Adobe's Generative AI Additional Terms, which govern their use of generative AI features in our services and software and explicitly require that users "must not remove or alter any watermarks or Content Authenticity Initiative metadata (e.g., Content Credentials) that may be generated with the output, or otherwise attempt to mislead others about the origin of the output."

Pictured: Content Credentials on an Image



Providing Resources

In 2023, the C2PA unveiled a new Content Credentials icon, a symbol of transparency designed to signal trustworthy digital content. After two years of collaborative research, design, and development, the community of technical leaders and UX experts unanimously agreed on this minimalist icon. It features the letters "CR" enclosed in a pin and can be etched into media like images and videos. When users scroll over the icon, it reveals a "digital nutrition label" with verified information about the content's origin, creation date, tools used, and any edits made. Major brands and industry leaders will integrate this icon to enhance digital content transparency from creation to consumption. The goal is for it to become universally recognized, akin to the copyright symbol, and restore trust and transparency online.



In 2023, the C2PA launched a new online hub called ContentCredentials.org as a central place to learn more about Content Credentials, as well as an ongoing resource for consumers and creators alike to go and verify content – <http://contentcredentials.org/>. This included a re-imaging of the “Verify” tool where anyone can go to learn more about the content they are consuming.

Supporting implementation:

With the mature, open C2PA standard and CAI free, open-source tooling, Content Credentials now spans a swiftly growing range of platforms and technologies, including cameras, smartphones, software and more. Below are just a few examples that shed light on Content Credentials momentum in the past year, including what we've achieved as part of the collective work from both the CAI and C2PA.

Leica introduces world's first camera with Content Credentials built-in

In October 2023, [Leica introduced](#) the world's first camera with Content Credentials built-in within the new Leica M11-P — revolutionizing the field of photojournalism by ensuring authenticity at the point of capture. As more photojournalists and creatives use this technology, consumers are going to start to see Content Credentials on more digital content — helping them better navigate the digital world.

Nikon plans to adopt Content Credentials into Camera models soon

Nikon has affirmed progress towards its commitment to bringing Content Credentials into upcoming camera models. With its vast set of consumer and professional customers worldwide, Nikon's future implementation will help empower broad awareness and adoption. Nikon is now [collaborating with international news agency, Agency France-Presse \(AFP\)](#), to begin practical verification of this image provenance function — specifically to support authenticity and reliability in photojournalism and the fact-checking process.

Qualcomm integrates Content Credentials at the chip level for next generation smartphones

Provenance technology will soon come to your smartphone with Content Credentials embedded at the chip level — supporting photo authenticity for consumers and their devices. In October 2023, Qualcomm [announced its latest Snapdragon 8 Gen3 mobile platform](#) that works with Truepic to support Content Credentials in camera systems, based on the global C2PA standard format

Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms.

In addition to the upward trend of progress explained in detail in Objective 1a, it is also worth noting the tangible progress in the two organisations which underpin the eco-systems' efforts on content provenance which Adobe has taken a very active role in.

The Content Authenticity Initiative (CAI) which was launched in late 2019 is an Adobe-led initiative with more than 3,000 members and partners working to increase trust online through provenance, which are the facts about the origins of a piece of digital content. In 2023, we reached over 2,000 members including — CEPIC, Dentsu, Omnicom Group, National Geographic Society, National Public Radio (NPR), Photoshelter, and Publicis Groupe. We currently have 78 CAI members in Australia, including Woolworths and the Australian Associated Press.

Internally, at Adobe we have a team of full-time employees dedicated to working on CAI. This includes engineers helping to develop and maintain our open-source tooling for the community, user-experience designers, and a team dedicated to recruiting partners, supporting adoption and growing the community globally.

In addition, Adobe is an active member of the independent standards organisation, the Coalition of Content Provenance and Authenticity, and sits on the Steering Committee which meets weekly, Chairs the Technical Working Group and has representatives on the Threats and Harms Task Force, plus support from Adobe employees from our Communications and Policy team for C2PA external engagement.

We are committed to working with other C2PA members such as Microsoft, BBC, ARM, Intel and Sony to ensure open technical specifications for provenance are maintained to the highest standards and used to implement content provenance across the eco-system in a manner that is interoperable and ultimately adopted by international standards organisations as the single, unified way to address disinformation by empowering users with transparency.

Objective 4: Empower consumers to make better informed choices of digital content.

As mentioned above, the mission of the work Adobe is leading in tackling mis/disinformation is focused on supporting the provision of ubiquitous tools to help consumers make better informed decisions about the content they are consuming online.

As more and more CAI members adopt provenance technology and creators and media publishers use the technology to disclose details about how their content is made and altered, we will have an increasing amount of provenance-enabled content available so that consumers can check the veracity of content.

To date, we know over 200M assets have been created with Content Credentials using Adobe products.

In 2023, we continued to help publish and improve on a "CAI Media Literacy Curriculum." We recognise that media literacy is a vital component of the fight to tackle mis/disinformation. These curricula, created in collaboration with education experts, are crafted to help students develop critical media and visual literacy skills to better navigate the ever-changing digital information landscape. Each curriculum includes a foundational unit as well as lessons for use in social studies, the arts, and English & language arts (ELA), with media literacy lessons and themes integrated throughout all components. In 2023, we also added lessons on Generative AI. These standards-aligned lessons introduce students to generative AI and engage them in critical conversations surrounding the technology.

These materials are available for free at Adobe Education Exchange - <https://edex.adobe.com/> a free education platform and will be available for use by educators and students globally.

As mentioned in 1a, we also have introduced a new "Content Credentials" icon, symbolizing transparency and trust in digital content. The initiative aims for widespread adoption, akin to the copyright symbol, to bolster online trust and transparency.

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.

In 2023, Adobe published research on EKILA, a new decentralized system designed to help artists and creators get proper recognition and payment for their work in generative AI. EKILA uses a sophisticated method to trace and credit the original sources of AI-generated images, aligning with the latest standards for tracking content origins (C2PA). Additionally, EKILA expands the use of non-fungible tokens (NFTs) by creating a new form of token that represents ownership rights, thus establishing a three-way link between the ownership, rights, and credit for an asset. This ORA framework allows creators to have control over how their work is used in AI training and to receive a share of the profits, such as royalty payments, when their creations are used in GenAI projects.

Objective 7: Signatories will publicise the measures they take to combat Disinformation.

Adobe has consistently published details of all the major milestones noted in Objective 1a. which link to Adobe blog pages. In addition, the CAI regularly posts blogs updating the community of progression of our mission and has recordings of the quarterly community webinars on the CAI website. CAI and Content Credentials were prominently featured on our large, annual [conferences such as Adobe MAX](#).

2023 saw a lot of prominent and public activity on Adobe's on Content Authenticity and Content Credentials. Our senior leaders have consistently talked about our role combating misinformation and development of content credentials. There are many examples, but one prominent [was an op-ed published](#) by our General Counsel talking the importance in combating misinformation in elections.

In 2023, we also held a CAI Symposium. This stands as a testament to the remarkable growth and impact of efforts around content provenance. The inaugural gathering in 2020 had 59 organizations, which focused on media literacy and the burgeoning challenges of misinformation. The symposium has burgeoned into a pivotal event, attracting over 200 participants this year. These attendees, ranging from industry leaders to policy experts, shared a unified dedication to fostering an open, trustworthy, and transparent digital ecosystem. The symposium showcased the significant strides made in deploying Content Credentials and the C2PA standard, highlighting innovations like the world's first Content Credentials-enabled camera and the integration of transparency measures into mobile devices. As we look towards the future, the CAI Symposium 2023 has laid a solid foundation for continued progress in ensuring the integrity of our digital media landscape. In addition, this year's conference held a working session to understand solutions for the growing problem of image abuse, where bad actors create deepfakes of people they know to cause shame or emotional harm.

In July 2023, our General Counsel and Chief Trust Officer Dana Rao testified before the US Senate Committee on Judiciary, Subcommittee on Intellectual Property. Dana Rao, Executive Vice President and General Counsel at Adobe, emphasizes the transformative power of AI in creativity and the need for responsible innovation. He highlighted Adobe's commitment to empowering creators through technologies like Content Credentials, which enable artists to attach a "Do Not Train" tag to their work, ensuring control over AI training data.

Adobe has enjoyed a strong relationship to the creative community have published many examples of how they are using Content Credentials. In 2023, we ran several artists and creative series that features how Content Credentials can help them secure trust and attribution in their work. [One featured Australian-born Melissa Findley.](#)

Concluding remarks

Addressing the issue of content authenticity at scale to tackle mis/disinformation is a long-term, interdisciplinary, collaborative mission. As demonstrated in this year's report, the work Adobe is undertaking with its own products, and the efforts we are leading with the CAI community and our active role in the C2PA is beginning to show real and meaningful progress. 2023 was a year of building momentum in utility and adoption, built upon the foundation of previous years.

Globally, we are constantly working to expand and diversify the membership of both CAI and C2PA and increase support among stakeholders for the standard. These efforts include encouraging software companies, device manufacturers, publishers, and social media platforms to adopt content provenance solutions to expose a wider range of consumers to these tools.

Specifically, in Australia, our goal over the next year is to continue to educate media outlets and journalists on the value of CAI so they can work with these tools embedded into their platforms, and to continue to socialize the C2PA standard with policy makers and stakeholders.

And this work is more essential than ever before with the arrival of mainstream generative AI. We are already seeing the implications of this new technology and how content authenticity can and will be a significant part of this new era of innovation. The democratization of use and scale of accessibility to these tools will have implications for how we tackle mis/disinformation.

Adobe

Australian Code of Practice on Disinformation and Misinformation

Apple Pty Limited: Apple News

2023 Annual Transparency Report
17 May 2024

Summary

The Apple News product is designed to promote quality journalism, with a focus on the quality of the content and its visibility. to prevent the propagation of Disinformation and Misinformation.

It is through this lens that Apple views its commitments under the Code.

Commitments under the Code

Apple has opted-in to the following commitments under the Code for its Apple News product:

1a: Reducing the risk of harms that may arise from propagation of disinformation and misinformation

1c: Allowing users to report content

1e: Recommender systems

2: Reducing advertising and/or monetisation incentives

4: Enabling users to make informed choices about source of content

6: Strengthening public understanding

7: Publishing its annual transparency report

Apple has not opted-in to the following commitments under the Code for its Apple News product as set out below:

1b: Informing users about managed or prohibited user behaviour

Apple News is a service for professional news-gathering organisations and publishers, not for the dissemination of user-generated content.

1d: Responses to reports about user behaviour

Apple News is a service for professional news-gathering organisations and publishers, not for the dissemination of user-generated content.

3: Reducing advertising and/or monetisation incentives

As Apple News does not provide users the ability to surface content to other users within News, this objective and its focus around inauthentic and other user behaviour leading to propagation of disinformation and misinformation does not apply to the Apple News product.

5: Improving awareness of the source of political advertising

Apple does not sell political advertising either directly or through its resellers.

7: Publicising measures in addition to its annual report

Although Apple has not opted into this commitment, Apple may voluntarily report on additional initiatives not otherwise referenced in this report on a case by case basis, as relevant.

Reporting against commitments

Outcome 1a: Reducing harm by adopting scalable measures

Apple News, with its human curation and vetting of publishers, has been designed to reduce the risk of harm that may arise from the propagation of disinformation and misinformation. Apple News includes a diverse range of publications, with differing perspectives on issues, recognising that comprehensive coverage of issues further reduces the risks of misinformation and disinformation.

An Apple editorial team evaluates outlets before they are onboarded on the platform. Outlets are evaluated to ensure they are credible, standards-based, professional organizations. Details on the process is accessible at <https://support.apple.com/guide/news-publisher/publishing-on-apple-news-apde42330c66/icloud>. Likely reasons for a publication not to be onboarded include where a publication publishes factual inaccuracies or fails to adhere to widely accepted journalistic standards. The onboarding process was further strengthened in mid 2022, with publishers no longer able to make an unsolicited application to join Apple News, so that the addition of new channels was on an invitation-only basis throughout 2023.

Apple News has worked with NewsGuard since 2020, meeting regularly to discuss potential new publishers on the platform as well as trends in misinformation and news narratives that may affect News. NewsGuard's ratings on credibility and transparency have been one of the many metrics we use to maintain a trusted, informative environment. In March 2023, NewsGuard expanded its service to Australia and New Zealand. This has given us access to local journalism experts and a more nuanced perspective on Apple News's presence within the greater Australia news environment and the misinformation narratives spreading there, including on climate change, the Voice to Parliament referendum and global conflicts.

The careful and deliberate human curation of stories by Apple News Editors on a range of topical issues is a key feature of the service. This includes a number of high visibility touchpoints that are entirely managed by our Apple News Editors:

- Top Stories section within Apple News and Apple Stocks featuring a selection of stories from across our portfolio of publishing partners
- Individually crafted Spotlight Collections where Editors select content, images and themes
- Weekly Newsletter to opted-in Apple News readers

Examples referenced above include:

Voice to Parliament referendum

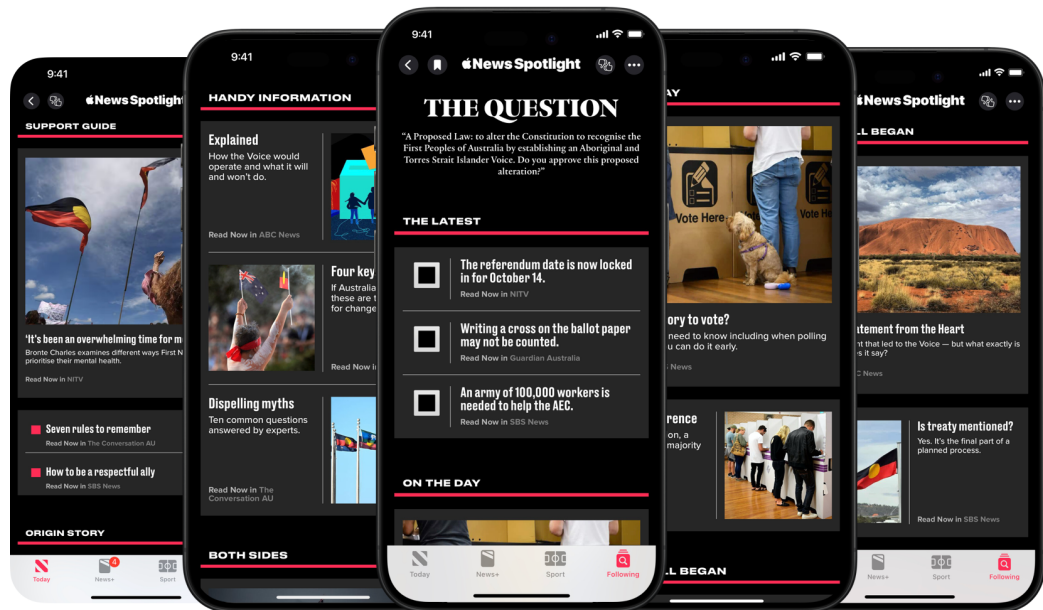
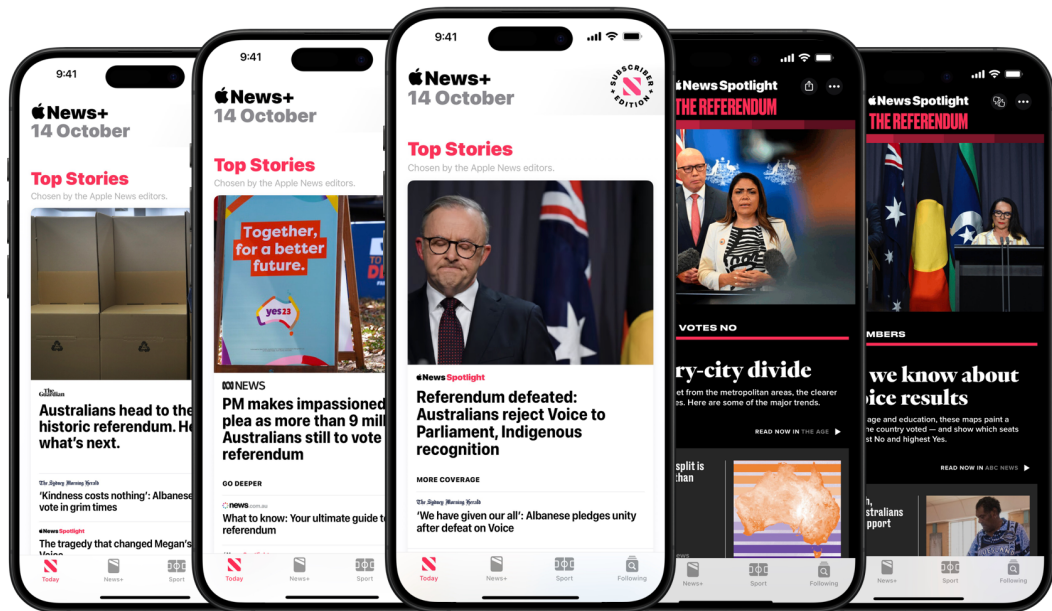
Apple News began its extensive coverage of the Voice in March 2023 in conjunction with the finalisation of the referendum question. Our coverage sought to navigate the nuance and complexity of the debate in an informed, thoughtful way for Apple News readers. We ensured all featured content was clear, fact-based and reflected the view of experts in the field, in conjunction with first-person perspectives to add depth and quality to a debate that often drove high levels of social fragmentation.

Between March and October, Apple News Editors crafted, featured and promoted 15 free Spotlight collections on the Voice, with over 200 individual pieces of content from a wide variety of quality publications. These collections received over 3 million views from readers. For many of these Spotlights, Apple News sent a push notification to all opted-in Apple News readers.

Our collections included timely insights, which addressed key moments and milestones during the debate, with the specific intent of mitigating misinformation in the community, including:

- Quality, well-sourced and authoritative journalism during debates around the constitutional implications of the referendum
- Fact-check pieces when the official "Yes" and "No" pamphlets were released

- Features on the origins and detail of the Uluru Statement from the Heart when this became a contentious issue.
- On the vote occurring, extensive coverage of the latest results and outcome, with follow-up Spotlights published on October 14 and 15.

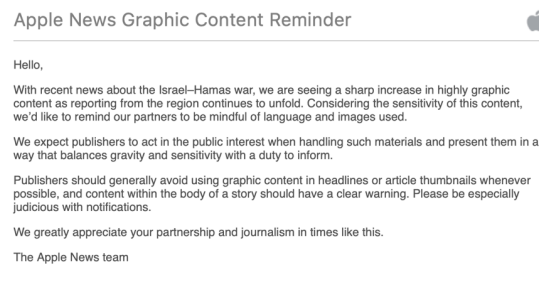


Israel-Hamas war

Apple News editors have sought to provide extensive coverage for Australian readers, drawing on the deep editorial expertise from our partner publications, with special emphasis on our international partners. Reuters, CNN, The Guardian and The Wall Street Journal were among many featured publications with journalists in the conflict zone, filing high-quality first person, well-researched reports. Our coverage has sought to inform Australians of the latest developments in the region and provide insights on the geopolitical situation in the Middle East, seeking to focus

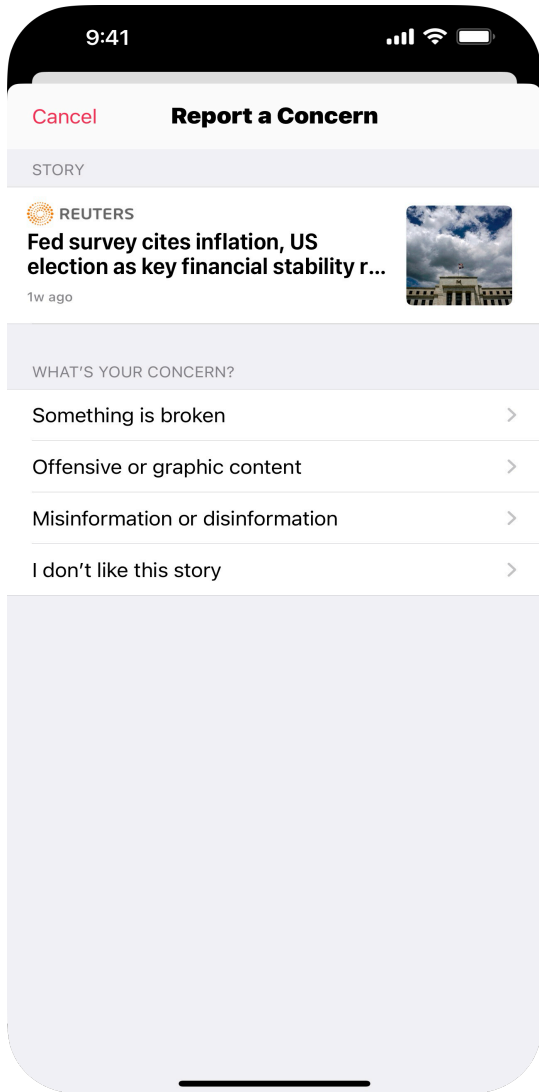
on verified and factual news reports at a time of heightened anxiety in the community.

Mindful of the impact of graphic content, and the potential for misuse of such content, we also took the opportunity to remind our publishers as to the appropriate use of graphic content.



Outcome 1c: Users can easily report offending content

Apple continues to enable customers to provide feedback on a per article basis (as per <https://support.apple.com/en-us/HT211226>), including permitting customers to specify the basis for their complaint (see screenshot below). Readers are prompted to choose from several categories of problems (the problems can be both technical and content-related, for example "Something is broken" or "Misinformation or disinformation"). A team of moderators then evaluates each report to determine whether the article violates the Apple News guidelines.



When the numbers of concerns raised by users are broken down, they demonstrate very limited misinformation/disinformation concerns being raised by our readers for articles made available by Australian news publishers, supporting our approach taken in the design of the Apple News product. In 2023, Apple News readers worldwide reported approximately 331,000 concerns on article content or with technical issues. The vast majority — approximately 326,000 — of these concerns were not deemed to be violations of platform guidelines. Approximately 4,800 concerns on 2,500 individual articles worldwide were deemed valid and warranted action from the moderation team, although these cover a range of issues and were not limited to misinformation/disinformation.

The following table shows a comparison with 2021 and 2022 numbers.*

	2021	2022	2023
Number of concerns reported	655,000	370,500	331,000
Number of concerns reported deemed valid	17,000 (on 5,600 individual articles)	6,500 (on 2,800 individual articles)	4,800 (on 2,500 individual articles)

As was the case for 2022, articles produced by Australian publishers that were actioned for misinformation/disinformation in 2023 accounted for less than one one-hundredth of one percent of total article views in the Australian Apple News app (with less than five articles in the Australian Apple News app requiring action for misinformation.).

When an article is actioned for containing misinformation or disinformation, a secondary team of trained journalists is involved in the evaluation to ensure the correct action is taken.

Outcome 1e: Recommender systems

Apple makes information about recommendations in News available to users, together with options and tools associated with those recommendations. See

<https://www.apple.com/legal/privacy/data/en/apple-news/>

Objective 2: Disrupt advertising and monetisation incentives for disinformation.

As set out in our previous reports, the design and structure of Apple News disrupts advertising opportunities for Misinformation/Disinformation by limiting its appearance on the platform in the first place. See also categories of advertisements not permitted to be made available in Apple News, including ads that are misleading or deceptive within the Advertising on Apple News Content Guidelines (<https://support.apple.com/guide/adguide/unacceptable-or-prohibited-content-guidelines-apd527d891a8/icloud>).

In addition, we provide warnings when we discover advertisements that violate the Content Guidelines and have the ability to block certain advertisers that repeatedly violate the Content Guidelines.

* There may be a number of reasons for the reduction in concerns raised with us. For example, many users report news articles that cover a topic that they dislike, rather than representing misinformation or other breaches of our guidelines. We also adjusted the location of reporting options in 2021 to improve the quality of user reports and minimise reports from users who had not read or engaged with the news content.

Objective 4: Empower consumers to make better informed choices of digital content.

As set out in our previous reports:

- Publishers are clearly identified on a per article basis, allowing users to determine the source of their news.
- We work with the news rating organisation, NewsGuard, and perform our own editorial evaluations to develop an understanding of all the publishers on the platform, and ensure that the most trustworthy sources are prioritise.
- Apple employs editors with newsroom experience in reputable Australian journalistic institutions to evaluate publishers on our platform, which will help ensure that reputable and trusted brands are surfaced to users, so as to help teach users to recognise credible sources of information.
- The most visible part of the News app is Top Stories (with approximately 15% of total article views coming from Top Stories), which features only fact-based journalism and is 100% curated by veteran journalists from the Australian news industry who vet each story for adherence to standard journalistic ethics. As such, Apple is establishing the credibility of certain publishers and brands within its ecosystem and helping train users to recognise credible sources of information by establishing trust with the brands regularly featured in Apple News.
- In the case of important topics of public interest, our 'hubs' (see 1a above) provide a broad range of content from a trusted range of publishers, allowing for diverse perspectives to be presented.

The delineation between News and Opinion content is an important journalistic distinction, ensuring readers are aware when they are being presented with an opinion piece as opposed to factual coverage of a story. Although outside the reporting period, a communication was sent to all Apple News Publishers on 8 January 2024 requesting clear opinion labelling be included in publisher headlines to reinforce this distinction. This communication also included a requirement that all content which is produced with the assistance of generative AI is appropriately labelled, with links to technical documentation for this requirement. Publishers were advised that "No AI generated content, including material from third party sources, should be published in Apple News without human editorial oversight."



Hi,

Apple News is committed to maintaining a trusted, informative news environment. To that end, we are implementing two new labeling requirements that will help our users better understand the content they see in Apple News.

Opinion content

We ask that all publishers label their opinion content published to Apple News directly in the headline. This will help readers understand that an article they see in their feeds is an opinion piece before they click in. This is reflective of what users experience on many of your owned and operated sites and aligns with industry standards.

Generative AI content

With the rapid advancement of generative AI tools, transparency about the use of AI for content creation purposes is critical. Content generated by or with the assistance of AI should be labeled as such, with a [byline or co-byline](#). Bylines should be clear, avoiding use of a broad "Staff" byline, for instance. AI generated or augmented images, video, or audio must clearly indicate that the content is not original and authentic. **No AI generated content, including material from third party sources, should be published in Apple News without human editorial oversight.**

AI may be used as a tool in idea generation, headline generation, research, or analysis without noting its use in an article if the article was composed by a human journalist.

To label AI use appropriately, AI generated content must include the "contentGenerationType" [metadata flag](#). Additionally, all AI generated content should include a clear note in the story explaining how AI tools were used.

Apple may update our policies over time. Please let us know if you have any questions.

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.

Apple has not received any research requests to date.

Apple has taken a global approach to supporting media literacy programs, with our continued support for The News Literacy Project (NLP). This organisation received support from Apple to advance its efforts in empowering young people with the critical thinking skills necessary in today's digital age, training the next generation on how to seek out accurate and reliable information amid an increasingly complicated news landscape.

As set out in our response for objective 1a, we consider our work with NewsGuard to also be relevant to this objective. Apple News has worked with NewsGuard, an organization that provides tools for readers to understand misinformation that may be spreading online, since 2020.

Objective 7: Signatories will publicise the measures they take to combat Disinformation.

As set out in our previous reports, Apple publishes this annual transparency report to outline measures taken.

Concluding remarks

Apple News is committed to creating a trusted, informative news environment by advancing quality journalism and, by design, to minimise the risks of misinformation and disinformation.

Australian Code of Practice on Disinformation and
Misinformation

Google Annual Transparency Report, May 2024

1st January 2023 - 31st December 2023

Introduction

As Google's mission is to organise the world's information to make it universally accessible and useful, combating misinformation and disinformation is of utmost importance.

Since Google's founding, our product, policy, and content enforcement decisions have been guided by the following three principles:

1. **Value openness and accessibility**
2. **Respect user choice**
3. **Build for everyone**

With these principles in mind, we implement a multi-faceted approach to address the complex challenges and risks raised by misinformation and disinformation across many of our products and services.

This report discusses these approaches across Google's diverse products and services, particularly in the context of addressing misinformation and disinformation. The contents of the report are organised through the framework of Objectives 1 through 7 of the Code, and cover the period from 1 January 2023 to 31 December 2023. We highlight relevant policies, product features and tools, and enforcement efforts across Google's key consumer-facing information services, such as Google Search, Google News, Google Advertising, and YouTube. We also include case studies on a range of initiatives including supporting the Australian Federal election campaign, and activities undertaken as a result of the Russia / Ukraine conflict and the Israel-Gaza conflict. Unless stated otherwise, all policies apply to users and content creators in Australia.

Some, but not all, of the updates that took place during 2023 are provided in this report for Google Advertising, Google Search and YouTube are listed below.

Google Advertising:

- Launched the [Ads Transparency Center](#), a searchable hub of all ads from verified advertisers, which helps people quickly and easily learn more about the ads they see on Search, YouTube and Display.
- Updated its Political Content Policy to require that all verified election advertisers in regions where verification is required (including Australia) must prominently disclose when their ads contain synthetic content that inauthentically depicts real or realistic-looking people or events.
- Inclusion of metrics on ad appeals in Australia for enforcement of misinformation policies.
- Processes are in place for advertisers who repeatedly violate certain Google Ads policies, including repeatedly placing digital advertisements that propagate Disinformation or Misinformation.
- Expanded Australian Election Ads policy to include ads that feature an Australian federal political party, a candidate for the Australian House of Representatives or Senate, a current elected

federal office holder in the Australian House of Representatives or Senate, or a referendum or plebiscite conducted by the Australian Election Commission. Additionally, this policy was expanded to include a state or territory political party, a candidate for elected state or territory office, a current elected state or territory officeholder, or a referendum or plebiscite conducted by a state or territory election commission.

Google Search

- Rolled out the 'About This Image' feature to users globally, giving people an easy way to check the credibility and context of images they see online and images' metadata.
- Launched a beta version of two new Fact Check Explorer features which allow a user to search an image for a fact-check and to see the full timeline of contextual information on an image.
- Included examples of how Google Search is providing authoritative information relating to the Israel-Gaza conflict.

YouTube

- Updated its Medical Misinformation Policy, removing content that promotes harmful or ineffective cancer treatments.
- Published its approach to responsible AI innovation, which includes its plan to require that creators disclose when they have created altered or synthetic content that is realistic, including using AI tools.
- Included new metrics on videos, channels and comments removed from YouTube as a response to combating misinformation as it relates to the Israel-Gaza conflict.

We will continue to publish updates to this report annually. We are also committed to improving and augmenting future iterations with further insights relevant to our continued efforts to combat misinformation and disinformation.

With respect to terminology, we acknowledge the differences between 'misinformation' and 'disinformation', as outlined by this Code. We use the term 'misinformation' to refer to both disinformation and misinformation and apply our policies and enforcement actions equally across each category.

Commitments under the Code

Google, including YouTube, has committed to all seven Objectives and related Outcomes provided in the Code and detailed below. As we respond and adapt to new and evolving challenges relating to misinformation we continually review our products, policies, enforcement and transparency work to ensure that they are as effective as possible.

Objective and Outcomes		Applicable service(s)
Objective 1 - Provide safeguards against harms that may arise from disinformation and misinformation		
1a	Signatories contribute to reducing the risk of harm that may arise from the propagation of disinformation and misinformation on digital platforms by adopting a range of scalable measures.	Google Search, Google News, Google Advertising ¹ , and YouTube
1b	Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.	
1c	Users can report content or behaviours to Signatories that violate their policies under section 5.10 through publicly available and accessible reporting tools.	
1d	Users will be able to access general information about Signatories' actions in response to reports made under 5.11.	
1e	Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.	
Objective 2 - Disrupt advertising and monetisation incentives for disinformation		
2	Advertising and/or monetisation incentives for disinformation and misinformation are reduced.	Google Advertising and YouTube
Objective 3 - Work to ensure the integrity and security of services and products delivered by digital platforms.		
3	The risk that Inauthentic User Behaviours undermine the integrity and security of services and products is reduced.	Google Search, Google News, Google Advertising, and YouTube
Objective 4 - Empower consumers to make better informed choices of digital content.		
4	Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify misinformation.	Google Search, Google Advertising, and YouTube
Objective 5 - Improve public awareness of the source of political advertising carried on digital		

¹ **Note:** Google Advertising refers to both Google Ads and Google AdSense where appropriate. Within the body of the report, 'Google Advertising' is used to refer to both services, whereas the individual service will be named if only applicable to Google Ads or Google AdSense.

platforms.		
5	Users are better informed about the source of political advertising.	Google Advertising
Objective 6 - Strengthen public understanding of disinformation and misinformation through support of strategic research.		
6	Signatories support the efforts of independent researchers to improve public understanding of disinformation and misinformation.	Google Search and YouTube
Objective 7 - Signatories will publicise the measures they take to combat disinformation and misinformation.		
7	The public can access information about the measures Signatories have taken to combat disinformation and misinformation.	Google Search, Google Advertising, and YouTube

Objective 1: Provide safeguards against harms that may arise from disinformation and misinformation.

Outcome 1a: Signatories contribute to reducing the risk of harms that may arise from the propagation of disinformation and misinformation on digital platforms by adopting a range of scalable measures.

The misinformation landscape, and the narratives that propagate through it, are constantly evolving. As such, our efforts and interventions to combat misinformation must adapt accordingly. Across many of our products and services, we deploy a range of measures to address the risk of potential harm caused by the propagation of misinformation. We also monitor certain narratives that pose a risk of harm, and may adjust our policies, and/or the enforcement of those policies, to counter them. We present a number of outcome-focused metrics to this effect below.

Actions taken to address coordinated influence operations

Our Threat Analysis Group (TAG) and Trust & Safety teams work to monitor malicious actors around the globe, disable their accounts, and remove the content that they post, including but not limited to coordinated influence operations and other operations that may affect Australia. TAG publishes a quarterly TAG [Bulletin](#), which provides updates about coordinated influence operation campaigns terminated on Google's platforms, as well as periodic [blog](#) posts detailing specific campaigns, threats, or trends.

Throughout 2023, TAG has identified several campaigns as part of their investigation into coordinated influence operations linked to Russia. In August 2023, TAG identified a campaign linked to the Internet Research Agency (IRA) that was sharing content in Russian that was supportive of Russia and critical of Ukraine, NATO and the United States. Additionally, TAG is closely monitoring activity in Israel and Gaza with regards to the ongoing conflict, focusing on the safety and security of users and the platforms that help them access and share important information.

Enforcement of YouTube's Community Guidelines

As detailed in our [Community Guidelines](#), YouTube does not allow misleading or deceptive content that poses a serious risk of egregious harm. We enforce our policies across the globe, including in Australia,

using a combination of content reviewers and machine learning to remove content that violates our policies as quickly as possible.

YouTube has long been updating, on a regular and ongoing basis, its internal systems and processes related to the detection of content that violates its policies. This includes investment in automated detection systems. In August 2023, YouTube announced changes to its [Community Guidelines warnings](#). After a creator's first Community Guidelines violation, they will typically get a warning with no penalty to their channel. They will now have the chance to take policy training to allow the warning to expire after 90 days. In November 2023, YouTube updated its [Community Posts Policy enforcement](#). While YouTube has always removed Community posts that violate its policies, starting on 17 November 2023, these posts may also result in a Community Guidelines strike applied to the channel. More information on how YouTube addresses misinformation can be found [here](#).

In 2023, the following actions were taken:

30,000,000+

YouTube **videos removed globally** for violating the Community Guidelines.

140,000+

YouTube **videos** that violated Community Guidelines and were uploaded from IP addresses in **Australia** were **removed**.

2,000,000+

YouTube **videos removed globally** for violating Misinformation or Spam, misleading and scams Policies.

20,000+

YouTube **videos** uploaded from IP addresses in **Australia** were **removed** for violating Misinformation or Spam, misleading and scams Policies.

More than 65% of videos that violated Community Guidelines and were uploaded from IP addresses in Australia were removed with 10 or fewer views.

As mentioned above, we rely on a combination of people and technology to flag inappropriate content and enforce these guidelines. These flags can come from our automated flagging systems, from members of the [Priority Flagger program](#) (previously known as the Trusted Flagger program) or from users in the broader YouTube community. YouTube's Priority Flagger program enables NGOs and government agencies who are particularly effective at identifying certain types of harmful content to have access to more sophisticated reporting processes and prioritised review of flags.

In addition to user flagging, YouTube uses smart detection technology to flag videos for review. YouTube developed powerful machine learning that detects content that may violate YouTube's policies and sends it for human review. In some cases, that same machine learning automatically takes an action, if there is high confidence that content is violative given information about similar or related content that has been previously removed.

Additionally, YouTube ensures integrity of its systems through:

- Having a dedicated team to identify and mitigate the impact of sophisticated bad actors on YouTube at scale, while protecting the broader community;
- Partnering with Google’s Threat Analysis Group (TAG) and Trust & Safety Teams to monitor malicious actors around the globe, disable their accounts, and remove the content that they post;
- Educating users about Community Guidelines violations through its [guided policy experience](#);
- Providing clear communication on [appeals processes and notifications](#), and regular [policy updates](#) on its Help Centre; and
- Investing in [automated systems](#) to provide efficient detection of content to be evaluated by human reviewers.

Where appropriate, YouTube makes it clear to users that it has taken action on their content and provides them the opportunity to appeal that decision.

In 2023, YouTube received **7,500+** appeals in Australia regarding a Community Guidelines violation removal decision. Following appeals from creators, **750+** removed videos were reinstated.

2023 Violative View Rate Estimate

YouTube strives to prevent content that violates our policies from being widely viewed — or viewed at all— before it is removed. As the overwhelming majority of violative content is detected by automated systems, YouTube’s Violative View Rate (VVR) is a good indication of how well our automated systems are protecting our community. VVR is an estimate of the proportion of video views that violate our Community Guidelines in a given quarter (excluding spam). In order to calculate VVR, we take a sample of the views on YouTube and send the sampled videos for review. Once we receive the decisions from reviewers about which videos in the sample are violative, we aggregate these decisions in order to arrive at our estimate.

- In **Q1** (Jan - Mar 2023), VVR was **0.08-0.10%** (i.e., out of every 10,000 views on YouTube, 8-10 were of violative content).
- In **Q2** (Apr - Jun 2023), VVR was **0.09-0.10%** (i.e., out of every 10,000 views on YouTube, 9-10 were of violative content).
- In **Q3** (Jul - Sep 2023), VVR was **0.10-0.11%** (i.e., out of every 10,000 views on YouTube, 10-11 were of violative content).
- In **Q4** (Oct - Dec 2023), VVR was **0.11-0.12%** (i.e., out of every 10,000 views on YouTube, 11-12 were of violative content).

Additional information about the VVR methodology is available in the YouTube Community Guidelines enforcement [transparency report](#) and a third-party [statistical assessment](#) commissioned by Google.

Case Study: YouTube's response to combating misinformation as it relates to Russia's invasion of Ukraine

YouTube's [Community Guidelines](#) are a key part of YouTube's [broader suite of policies](#) and are [regularly updated](#) in consultation with outside experts and YouTube creators to keep pace with emerging challenges or crises. YouTube's teams work quickly to remove content that violates its policies.

From 24 February 2022 to 31 December 2023, YouTube removed more than 12,000 channels and more than 140,000 videos related to the ongoing crisis in Ukraine for violating its content policies, including those pertaining to misinformation, hate speech, and graphic violence.

Case Study: YouTube's response to combating misinformation as it relates to the Israel / Gaza conflict

Following the terrorist attack by Hamas in Israel and the escalated conflict now underway in Israel and Gaza, YouTube removed over 95,000 videos, terminated over 4,500 channels and removed over 70 million comments as of 8 January 2024.

YouTube's [Community Guidelines](#) are a key part of YouTube's [broader suite of policies](#) that are [regularly updated](#). One example is YouTube's [Hate Speech Policy](#) which prohibits content denying, trivialising, or minimising violent historical events, including the 7 October Hamas attacks in Israel.

From 8 October 2023 to 26 October 2023, YouTube launched a crisis resource panel to highlight authoritative and verified information from Israeli authorities for users in Israel. The crisis resource panel directed users towards resources like the Israeli National Emergency Portal.

Streamlining Medical Misinformation Guidelines for Transparency

On 15 August 2023, YouTube provided an update on its long term vision for YouTube's medical misinformation policies. As medical information – and misinformation – continuously evolves, YouTube needs a policy framework that endures over the long term, and preserves the important balance of removing egregiously harmful content while ensuring space for debate and discussion. While specific medical guidance can change over time as we learn more, YouTube's goal is to ensure that when it comes to areas of well-studied scientific consensus, YouTube is not a platform for distributing information that could harm people. Moving forward, YouTube will streamline dozens of our existing [medical misinformation guidelines](#) to fall under three categories – Prevention, Treatment, and Denial. These policies will apply to specific health conditions, treatments, and substances where content contradicts local health authorities or the World Health Organization (WHO). YouTube's approach is to be clear and transparent, so that content creators understand where the policy lines are, and viewers know they can trust the health information they find on YouTube.

Actions taken to combat AI generated mis/disinformation

Google Search 'About this image'

In October 2023, Google Search [rolled out the 'About this image'](#) feature to English language users globally. 'About this image' gives people an easy way to check the credibility and context of images they see online, including an image's history, how other sites use and describe it, and an image's metadata.

With added insights from 'About this image', users will know if an image may have been generated with Google's AI tools when they come across it in Search or Chrome. All images generated with Imagen 2 in Google's consumer products will be marked by SynthID, a tool developed by Google DeepMind that adds a digital watermark directly into the pixels of images generated. SynthID watermarks are imperceptible to the human eye but detectable for identification. Google Search hopes its [SynthID technology](#) can work together with a broad range of solutions for creators and users across society, and it is continuing to evolve SynthID by gathering feedback from users, enhancing its capabilities, and exploring new features.

SynthID could be expanded for use across other AI models and Google Search is exploring the potential of integrating it into more Google products and making it available to third parties in the near future — empowering people and organisations to responsibly work with AI-generated content.

YouTube's approach to responsible AI innovation

In November 2023, YouTube published its [approach to responsible AI innovation](#), which outlined its plan to require that creators disclose when the content they are uploading is made with altered or synthetic media, including generative AI, and is realistic - that is, a viewer could easily mistake what is being shown with a real person, place or event. YouTube announced that, tied to these disclosures, it will begin applying labels to content indicating that some of the content was altered or synthetic, as well as a more prominent label for certain types of content about sensitive topics. YouTube officially rolled out the disclosure requirements and tools for creators in Q1 2024 i.e. after the reporting period for this report.²

YouTube also announced plans to make it possible to request the removal of AI-generated or other synthetic or altered content that simulates an identifiable individual, including their face or voice, using its [privacy request](#) process. Not all content will be removed from YouTube, and YouTube will consider a variety of factors when evaluating these requests. This could include whether the content is parody or satire, whether the person making the request can be uniquely identified, or whether it features a public official or well-known individual, in which case there may be a higher bar.

YouTube's guiding approach is to ensure we are providing transparency to users in a way that enables creators to experiment with and harness the potential of AI. To build responsibility into its AI tools and features, YouTube is investing in significant, ongoing work to develop guardrails to safely deploy AI tools for YouTube creators. YouTube will also incorporate user feedback and learning to continuously improve protections. Within YouTube, dedicated teams such as the intelligence desk are specifically focused on

² These have been rolled out in Q1 2024. More details can be found in YouTube's blogpost on [how we're helping creators disclose altered or synthetic content](#).

adversarial testing and threat detection to ensure YouTube's systems meet new challenges as they emerge.

Google Ads Update to Political Content policy

In mid-November 2023, Google Ads [updated its Political content policy](#) to include disclosure requirements for synthetic content. This policy requires that all verified election advertisers in regions where verification is required must prominently disclose when their ads contain synthetic content that inauthentically depicts real or realistic-looking people or events. This disclosure must be clear and conspicuous, and must be placed in a location where it is likely to be noticed by users. This policy will apply to image, video, and audio content.

Examples of ad content that would require a clear and conspicuous disclosure include (non-exhaustive):

- An ad with synthetic content that makes it appear as if a person is saying or doing something they didn't say or do; and
- An ad with synthetic content that alters footage of a real event or generates a realistic portrayal of an event to depict scenes that did not actually take place.

Acceptable disclosure language will vary depending on the specific context of the ad, but some examples may include:

- *This audio was computer generated.*
- *This image does not depict real events.*
- *This video content was synthetically generated.*

Outcome 1b: Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.

We aim to ensure that our policies across products and services are available to the public, users and creators in a form that is clear, predictable and repeatable. Each of our product policies address the types of behaviours and content prohibited on the product, with examples as needed. A list of product-specific policies are available on [this page](#). Those most relevant for this report are included in [Appendix A](#), along with explanations of each policy.

Outcome 1c: Users can report content or behaviours to Signatories that violates their policies under section 5.10 through publicly available and accessible reporting tools.

Our products and services provide publicly available and accessible channels or mechanisms that allow users to report content that they believe has violated our policies. Highlighted below are examples of

reporting mechanisms that users can utilise to provide feedback across Google Search, Google Ads, Google AdSense and YouTube.

- In **Google Search**, users can provide feedback on an overall Search results page or on specific features such as [Knowledge Panels](#) or [Featured Snippets](#).
- On Google-served **Ads**, users can tap the three dots on the top right corner of an ad and select 'Report this Ad' to let us know about ads that they believe are illegal or violate our policies. Clicking on 'Report this Ad' will redirect users to the 'Report an ad/listing' [form](#). Using this feature, individuals can choose the reason that best describes their complaint.
- On **Google AdSense**, users can [report a site](#) that they believe is showing ads in violation of our product policies.
- On **YouTube**, users can [report](#) content using YouTube's flagging feature and indicate information about which of our policies they believe the video is violating. Users can also report inappropriate channels, playlists, comments and other content.

Outcome 1d: Users will be able to access general information about Signatories' actions in response to reports made under 5.11.

We provide regular, publicly available reports on enforcement of our content policies - these include information regarding actions in response to reports of misinformation. A list of these reports (including public links to the materials) is included below. Information from these sources have also been incorporated throughout the relevant sections of this report.

- Our [Google Transparency Report website](#) is a centralised hub for transparency reporting on key content topics across various Google products and services;
- The annual [Ads Safety Report](#) provides updates on policy enforcement in Google Ads;
- Our Threat Analysis Group Quarterly Bulletin (published on our [Threat Analysis Group blog](#)) discloses actions we have taken against coordinated influence operation campaigns on our platforms;
- Our [YouTube Community Guidelines Enforcement report](#) provides a quarterly update on the work we do to enforce our policies on YouTube.

Outcome 1e: Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.

Google's ranking and recommender systems

Google has long invested in ranking and recommender systems that seek to connect people with authoritative sources and elevate high quality content. These systems enable users to make informed choices about content encountered, as further described in Outcome 4 (with additional tools and features highlighted in Appendix B). The below sub-sections highlight some of the ranking and recommender systems in place, as well as user options to manage their recommendations.

Ranking Google Search results

We continue to improve the design of our ranking systems, which are key to helping users make informed decisions and reducing the proliferation of misinformation. For example, Google Search's algorithms consider a host of 'signals' (or characteristics of a web page) that are indicative of high-quality and reliable information and undergo a [rigorous testing process](#) that involves both live tests and thousands of trained external Search Quality Raters from around the world. Raters do not determine the ranking of an individual, specific page or website, but they help us benchmark the quality of our results so that we can meet a high bar for users of Google Search all around the world.

Under our [Search Quality Rater Guidelines](#), raters are instructed to assign the lowest rating to pages that are potentially harmful to users or specified groups, misleading, untrustworthy, and spammy. Examples of such pages include those that contain clearly inaccurate harmful information that can easily be refuted by straightforward and widely accepted facts; harmful information that contradicts well-established expert consensus; and harmful, unsubstantiated theories/claims not grounded in any reasonable facts or evidence.

Information such as user location, past Google Search history, and [Search settings](#) all help Google Search ensure user results are what is most useful and relevant at that moment. The user can control what Google Search activity is used to improve their experience, including adjusting what data is saved to their Google account at [myaccount.google.com](#). To disable Google Search personalisation based on activity in a user's account, the user can [turn off Web & App Activity](#). Additionally, users are able to decide whether Search shows personal results based on the information provided in their Google account by managing the 'Show personal results setting'. Users also have the option to [browse the web privately](#) in Incognito mode if they do not want Google Chrome to remember their activity. Google Search systems are designed to match a user's interests, but they are not designed to infer sensitive characteristics like race, religion or political party.

YouTube recommendation systems

YouTube has expanded the ways in which it ensures that its ranking and recommendations systems surface high quality content to curb the spread of harmful misinformation and 'borderline' content — content that comes close to, but does not quite violate YouTube's Community Guidelines. An explanation of how our recommendation system works, including how we raise up authoritative information in ranking and recommendations, can be found [here](#).

Signals used to recommend content

YouTube's recommendation system is constantly evolving, learning every day from information that YouTube calls signals, including but not limited to: watch history, search history, channel subscriptions, likes, dislikes, and satisfaction surveys. There are several ways users can influence their recommendations and search results. Users can remove specific videos from their [watch history](#) and searches from their [search history](#). Users can also pause their watch and search history, or start fresh by clearing their watch and search history. Additional information about how a user can manage their recommendation settings are outlined [here](#) in YouTube's Help Centre.

User choice and control on ads

Google is committed to giving users transparency, choice and control when it comes to the ads they see on Google platforms. That is why Google has long offered tools like Ad Settings which allows people to control how ads are personalised or even opt out of personalised ads altogether, as well as features like 'Why this ad?' which helps explain why a specific ad is being shown. [My Ad Center](#) offers tools that let users choose the kinds of ads they see on Google and YouTube. These tools are only available when personalised ads are on.

Objective 2: Disrupt advertising and monetisation incentives for disinformation.

Outcome 2: Advertising and/or monetisation incentives for disinformation and misinformation are reduced.

Relevant policies across our products and services

Our advertising and monetisation policies prohibit a range of behaviours and types of content that are clearly connected to misinformation, or that commonly overlap with misinformation. These policies include, but are not limited to:

- [Google Ads Policies](#)
- [Google Publisher Policies](#)
- [YouTube Advertiser-friendly Content Guidelines](#)
- [YouTube Channel Monetisation Policies](#)

Further details on these policies can be found in [Appendix A](#).

Updating our monetisation policies related to misinformation

As the misinformation landscape and the narratives that propagate through it constantly evolve, our efforts and interventions to combat misinformation must adapt accordingly. This includes updating our policies and monitoring risks associated with misinformation in the context of broad societal issues that impact our users' lives. Examples of updates made to our Ads and YouTube policies are explained below.

Google Ads also provides its advertising partners with features that enable them to maintain control over where their ads appear, the format in which their ads run, and their intended audience. In addition to performing regular review of our monetisation policies, Google Ads leverages its Dynamic Exclusion Lists feature to help our advertisers to seamlessly and continuously prevent ads from serving alongside certain content. Further information can be found [here](#).

Enforcing our policies to reduce monetisation incentives for misinformation

To verify that advertisers and publishers on our networks are complying with our policies, we continuously monitor our advertising networks and use a combination of algorithmic and human reviews. The metrics below highlight Australia-specific enforcement actions taken for violation of misrepresentation-related ads policies. Google AdSense provides a way for publishers to earn money

from their online content. AdSense works by matching ads to publisher sites based on site content and visitors. The ads are created and paid for by advertisers who want to promote their products.

647,703

AdSense pages actioned for violating Unreliable and Harmful Claims, Replicated Content, Manipulated Media, or Dangerous or Derogatory Content, Deceptive Practices, and Shocking Content Policies in Australia.

136

AdSense domains actioned for violating Unreliable and Harmful Claims, Replicated Content, Manipulated Media, or Dangerous or Derogatory Content, Deceptive Practices, and Shocking Content Policies in Australia.

35,392,987

creatives actioned for violating [Destination Requirements Policies](#) (i.e. Insufficient Original Content) in Australia.

162,140

creatives actioned for violating [Inappropriate Content Policies](#) (i.e. Dangerous or Derogatory Content, Shocking Content, Sensitive Events, Animal Cruelty) in Australia.

762,078

creatives actioned for violating [Misrepresentation Policies](#) (i.e., Unacceptable Business Practices, Coordinated Deceptive Practices, Misleading Representation, Manipulated Media, Unreliable Claims, Misleading Ad Design, Clickbait Ads, Unclear Relevance, Unavailable Offers) in Australia.

Ads that do not follow Google Ads policies will be disapproved or (if appropriate) limited in where and when they can show. Advertisers have multiple options and pathways to appeal a policy decision directly from their Google Ads account.

In 2023, Google Ads received 50,461 ads appeals in Australia for the following policies: [Destination Requirements](#), [Inappropriate Content](#), and [Misrepresentation](#). Of these, 13,579 were successful (i.e., all entities that were reviewed as part of the appeal were overturned), 5,979 were partially successful (i.e., some entities were overturned) and 30,903 failed (i.e., no entities were overturned, as we confirmed they were correctly labelled initially).

For more information about the appeal process, check the [Help Centre page](#).

Our [2023 Ads Safety Report](#) and [Ads Safety blog post for 2023](#) are publicly available and contain additional data that exemplify enforcement actions taken on ads and publisher content.

Efforts to deter advertisers from repeatedly placing digital advertisements that propagate Disinformation or Misinformation

Google Ads prohibits [coordinated deceptive practices](#) on its platform and takes violations of this policy very seriously. If we find violations of this policy, we will suspend the advertisers' accounts upon detection and without prior warning, and they will not be allowed to advertise with Google Ads again. For more information, see [here](#).

Furthermore, Google Ads [updated its misrepresentation policy](#) to better enable us to rapidly suspend the accounts of advertisers who entice users to part with money or information by impersonating or falsely implying affiliation with or endorsement by a public figure, brand, or organisation. We've trained our automated enforcement models to detect these ads and begin removing them at scale.

Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms.

Outcome 3: The risk that Inauthentic User Behaviours undermine the integrity and security of services and products is reduced.

Google is continually working to address and mitigate risks associated with behaviours that seek to undermine the integrity and security of our products and services ('Inauthentic User Behaviours'). Targeted policies are in place across our products and services, tailored to the specific risks faced by each product or service in relation to Inauthentic User Behaviour. Further details on these policies can be found in [Appendix A](#).

Google Search & Google News

- Google Search [Webmaster Guidelines](#) prohibit techniques which may be misused to deceive our ranking systems or users.
- The [Google Search Content Policies](#) include policies related to Search Features. Content on Google News must follow all Google Search Content Policies, as well as [Google News-specific Policies](#).

Google Ads

- Google Ads [Advertising Policies](#) list unacceptable practices, content, and behaviours that advertisers must avoid.

Google AdSense

- AdSense users who wish to monetise their content with Google ad code are required to adhere to the [AdSense Program Policies](#).

YouTube

- Our [YouTube Community Guidelines](#) include policies to prohibit content intended to impersonate a person or channel, as well as fake engagement, which aims to artificially increase the number of views, likes, comments, or other metrics either by using automatic systems or serving up videos to unsuspecting viewers.

- More information related to 'Inauthentic User Behaviours' can be found in the Threat Analysis Group case study (see [Outcome 1a](#)) and advertisement policy enforcement metrics (see [Outcome 2](#)).

Objective 4: Empower consumers to make better informed choices of digital content.

Outcome 4: Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify misinformation.

Tools and features enabling users to make more informed choices

Across many of our products, we provide users with a variety of opportunities to make informed choices about content encountered, thereby allowing users to identify misinformation. With more tools to identify misinformation, users are then empowered to report instances of misinformation using the mechanisms outlined in [Outcome 1c](#), thus helping reduce the overall risk of harm from misinformation.

Case Study: How Google supported the 2023 Aboriginal and Torres Strait Islander Voice Referendum

In October 2023, Australians voted in a national referendum on a proposal to enshrine an Aboriginal and Torres Strait Islander Voice in the Australian Constitution. On top of [measures already in place](#), Google provided support to help people participate and stay informed, protect vote integrity, and assist Yes and No campaigns to manage their digital presence. Some ways Google provided further support include:

- **Connecting voters to helpful and authoritative information** by working with the Australian Electoral Commission to surface official information about enrolment and voting on Google's platforms (including Search and YouTube). This included information about the referendum and how to vote, and prompts to go and vote on referendum day. YouTube also ran a prominent banner in the lead up to the referendum that linked Australian users to an Australian Electoral Commission playlist of informative videos.
- **Helping voters better understand the political advertising they see** by [expanding the Political Content policy](#) in May 2023. This requires advertiser verification, in-ad disclosures, and that any ad which features any of the following be included in our [political advertising transparency report](#):
 - a. An Australian federal political party, a candidate for the Australian House of Representatives or Senate, a current elected federal officeholder in the Australian House of Representatives or Senate, or a referendum or plebiscite conducted by the Australian Electoral Commission;

- b.** A state or territory political party, a candidate for elected state or territory office, a current elected state or territory officeholder, or a referendum or plebiscite conducted by a state or territory election commission, from any of the following states and territories: Australian Capital Territory, New South Wales, Northern Territory, Queensland, South Australia, Tasmania, Victoria and Western Australia.
- **Protecting information online** by (i) working with the different Voice campaign groups, electoral bodies and civil society groups to help everyone understand digital best practices and their responsibilities through [Google Ad policies](#) and [YouTube Community Guidelines](#) and (ii) raising up authoritative sources and removing violative content quickly using a combination of machine learning and people.
- **Collaborating with Australian Associated Press (AAP)** via the [Google News Lab](#) to provide and distribute fact-checks to the 300 Australian news publishers that subscribe to its service. Helping to ensure Indigenous journalists' experiences are included in reporting has also guided Google's partnership with AAP, which is why Google supported the placement of two Indigenous trainees in their newsrooms.

See Google's blog post on how it supported the 2023 Aboriginal and Torres Strait Islander Voice Referendum [here](#). More information about how Google is supporting democratic processes around the world is available [here](#).

The below sub-sections highlight some of the tools and features created to elevate authoritative sources and to help users make informed choices; an overview of additional tools and features for these purposes can be found in [Appendix B](#). Note that these tools and features are automatically available to all users in Australia, and their availability does not require an individual user to select and/or activate them.

Google Search content advisories

Google Search's content advisory notices help alert users to when they have encountered a query and results set that may not yet include high quality information from reliable sources or when the results retrieved are likely to be off-topic and therefore unhelpful. These are specifically designed to address data voids which include queries for which either content is limited or nonexistent or when a topic is rapidly evolving and reliable information is not yet available for that topic.

Google Search releases these content advisories, following both user research and multiple rounds of consultations with academic experts in mis- and disinformation. Content advisories include:

- those when a topic is rapidly evolving, available to users globally (see more regarding this feature [here](#)); and
- where Google Search systems do not have high confidence in the overall quality of search results in English (see [blog](#) on New Ways We're Helping you Find Quality Information and this [blog](#) for further details).

YouTube information panels

YouTube highlights information from authoritative third-party sources using information panels. As users navigate YouTube, they might see a variety of different information panels, including but not limited to:

- **Panels on topics prone to misinformation:** Topics that are prone to misinformation, such as the moon landing, may display an information panel at the top of search results or under a video. These information panels show basic background info, sourced from independent, third-party partners, to give more context on a topic. The panels also link to the third-party partner's website. YouTube continues to expand its deployment of these information panels globally. More details can be found [here](#).
- **Information panel providing publisher context:** If a channel is owned by a news publisher that is funded by a government, or publicly funded, an information panel providing publisher context may be displayed on the watch page of the videos on its channel. The information panel providing publisher context explains how the publisher is funded and provides a link to the publisher's Wikipedia page. More details can be found [here](#).

Information panels provide additional context, with each designed to help users make their own decisions about the content they find. These information panels appear in relevant search results and video watch pages, regardless of what opinions or perspectives are expressed. In 2023, YouTube displayed information panels on topics prone to misinformation below relevant videos or above search results **over 80 million times** in Australia.

Google Search structured authoritative information

Through '**SOS Alerts**', Google brings together relevant and authoritative content from the web, media, and Google products, and then highlights that structured content across Google products such as Google Search and Google Maps. The content includes authoritative help links and relevant local information aiming to make emergency information more accessible during a crisis. See [Help Centre](#) for more information. In 2023, there were **over 4,200,000 views/impressions** on Crisis Response alerts (e.g., 'SOS Alerts', 'Public Alerts') in Australia.

Case Study: Providing authoritative information on Search during the Israel-Gaza conflict

- Google Search launched an SOS Alert on 7 October 2023 to highlight authoritative and verified information related to the terrorist attacks and the ongoing hostage situation for users in Israel. The SOS Alert directed users to governmental and expert resources including the Home Front Command, blood donation, missing persons hotlines, and mental health and trauma support. The SOS Alert ran until 26 October 2023.
- Google Maps & Local Search added locations of authoritative bomb shelters for users in Israel. It also added a warning to listings of certain critical businesses in Israel and Gaza to alert users that the listed hours may no longer be accurate due to the conflict.

Media literacy

In the face of near limitless access to information, Google remains committed to supporting efforts that deepen users' collective understanding of misinformation. Google aims to improve users' media literacy and empower users to think critically through investing in media literacy campaigns and designing tools and features in a way that allows users to feel confident and in control of the information they consume and the choices they make.

Google Search

Google Search aims to connect users with high quality information, and help users understand and evaluate that information. Google Search has deeply invested in both information quality and information literacy, as described below.

Google Search's 'About This Result' feature enables users to quickly learn more about a result and make a more informed decision about the sites they may want to visit and what results will be most useful for them (more information found [here](#)). Further information on this feature is available in [Google's 2022 Australian Code of Practice on Disinformation and Misinformation Annual Transparency Report](#). In 2023, the 'About this Result' panel was **viewed 59,758,626 times** in Australia.

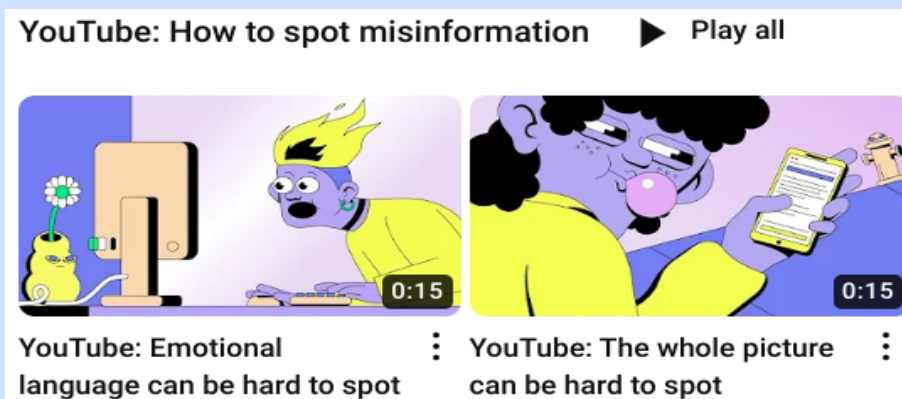
The 'More About This Page' link within the 'About This Result' feature provides additional insights about sources and topics users find on Google Search. When a user taps the three dots on any search result, they will be able to learn more about the page. Additional information can be found in the Google Search blog post [here](#). In 2023, the 'More About This Page' feature was **viewed 1,765,504 times** in Australia.

As mentioned in Outcome 1a, Google Search [rolled out the 'About this image'](#) feature to English language users globally in October 2023. 'About this image' gives people an easy way to check the credibility and context of images they see online, including an image's history, how other sites use and describe it, and an image's metadata.

YouTube

YouTube takes its responsibility efforts seriously, outlining clear [policies](#) used to moderate content on the platform and providing content that viewers can leverage to improve their digital media literacy skills. To empower viewers to think critically and share responsibly, YouTube invests in media literacy campaigns to support educating viewers on how to spot misinformation online; the most recent global media literacy campaign is highlighted below.

Case Study: YouTube's 'Hit Pause' Program



Source: <https://www.youtube.com/@HitPause>

YouTube's global media literacy program, 'Hit Pause' was developed in partnership with The National Association for Media Literacy Education (NAMLE). The program launched in November 2022 and is available in over 70 countries around the globe, including Australia where the program is co-branded with the Alannah and Madeline Foundation.

'Hit Pause' seeks to teach viewers critical media literacy skills through engaging and educational public service announcements (PSAs) via YouTube home feed and pre-roll ads, and on a dedicated [YouTube playlist](#). Throughout 2023, the program continued to roll out new videos that provided members of the YouTube community with the opportunity to increase critical thinking skills around identifying different manipulation tactics used to spread misinformation – from using emotional language to cherry picking information.

In 2023, there were **over 45 million** impressions of YouTube's 'Hit Pause' campaign in Australia.

Fact checks on Google Search and News

Fact checking is an important part of tackling misinformation. Google Search has developed policies and processes to label fact-checked articles that are displayed in Search results (if they meet the eligibility and technical criteria). Google also provides tools like [Fact Check Explorer](#) and the [Google FactCheck Claim Search API](#). Fact Check Explorer allows anyone to explore the Fact Check articles that use the '[ClaimReview](#)' HTML mark-up, an open standard that lets any fact-checker mark up their content so it can be identified and used by any online service (including but not limited to Search Engines or Social Media). To make it easier for fact-checkers to leverage the ClaimReview mark-up, Google developed a [free tool](#) that simplifies the process of marking up webpages using this standard. Additional information about the Fact Check Markup Tool can be found [here](#). Using the [Google FactCheck Claim Search API](#), users can query the same set of Fact Check results available via the [Fact Check Explorer](#) or a developer could continuously get the latest updates on a particular query. Fact-checks from the Australian Associated Press (AAP) are included in this database. The metrics below relate to the availability of fact-checked articles and use of the Fact Check Explorer tool.

73,093

articles available in English in Google Search **Fact Check Explorer** at the beginning of 2023, globally.

115,472

articles available in English in Google Search **Fact Check Explorer** at the end of 2023, globally.

4,904

Fact Check Explorer tool users in Australia in 2023.

Fact-check features on Google are another way to easily find information that has been verified by independent fact-checking organisations. The 'Fact Check' label in Google Search applies to published stories with fact-checked content that is indicated by the [schema.org ClaimReview markup](#), like round-up stories that contain multiple fact-check analyses within a single article. Google News may apply this label to publisher content, where applicable. This helps users find fact-checked content for major stories appearing on Google News; when browsing Google News on desktop, they can see recently fact-checked claims from independent publishers in their region, when related to the top stories of the day. Likewise, when users search for a topic that may be disputed, they might see fact-check articles in their results. These results display snippets to help users get context about a claim that was made.

In June 2023, Google Search launched a beta version of two new Fact Check Explorer features which allow a user to (i) search by image to see if a fact-check has been written on it already; and (ii) see the context and timeline of an image to see when it was first indexed by Google and how it has been used since. More information is available [here](#).

In October 2023, Google Search [announced](#) a beta version of Image Search functionality in the FactCheck Claim Search API under which approved journalists and fact-checkers will be able to search the fact-check image corpus on Fact Check Explorer via an API and integrate the knowledge into their own solutions. This will make it even easier for them to investigate images and build unique products for their readers.

More information about fact-check features can be found on '[Check the facts with these Google features](#)'.

Ads Transparency Center

We want to empower users to make informed decisions about the ads and advertisers they see through Google. This means providing greater transparency about who our advertisers are, where they are located, and which ads they show. This is why, in 2023, Google Ads launched the [Ads Transparency Center](#), a searchable hub of all ads from verified advertisers, which helps people quickly and easily learn more about the ads they see on Search, YouTube and Display. This is part of Google's wider efforts to provide users transparency, choice, and control in the ads they see.

Objective 5: Improve public awareness of the source of political advertising carried on digital platforms.

Outcome 5: Users are better informed about the source of political advertising.

Google's broader ads policies, as described in [Appendix A](#), apply to all ads, including election ads. Australia Election Ads are ads that feature: an Australian federal political party, a candidate for the Australian House of Representatives or Senate, or a current elected federal office holder in the Australia House of Representatives or Senate. Ahead of the referendum on an Aboriginal and Torres Strait Islander Voice, we updated the definition of Australia Election Ads to include ads that feature any of the following:

- An Australian federal political party, a candidate for the Australian House of Representatives or Senate, a current elected federal office holder in the Australian House of Representatives or Senate, or a referendum or plebiscite conducted by the Australian Election Commission;
- A state or territory political party, a candidate for elected state or territory office, a current elected state or territory officeholder, or a referendum or plebiscite conducted by a state or territory election commission, from any of the following states and territories: New South Wales, Victoria, Queensland, South Australia, Western Australia, Tasmania, Northern Territory, Australian Capital Territory.

Additionally, in mid-November 2023, Google Ads [updated its Political Content Policy](#) to require that all verified election advertisers in regions where verification is required (including Australia) must prominently disclose when their ads contain synthetic content that inauthentically depicts real or realistic-looking people or events. This disclosure must be clear and conspicuous, and must be placed in a location where it is likely to be noticed by users. This includes:

- Ads with synthetic content that makes it appear as if a person is saying or doing something they did not say or do;
- Ads with synthetic content that alters footage of a real event or generates a realistic portrayal of an event to depict scenes that did not actually take place.

Political Advertising Transparency Report - User interface & tools

Google has made significant investments in [enhancing transparency around election advertising](#). In addition to election ads verification policies, the [AU Political Advertising Transparency Report](#) makes it easy for voters, researchers, and journalists to see - among other things - who is purchasing election ads on Google, YouTube, and Partner properties in Australia and how much money is being spent on those ads.

Anyone can access and use this information, and Google has aimed to ensure that it is easy for third parties to analyse it more effectively: the report is searchable and downloadable, and can be filtered by spend, number of impressions, type of ad format, time and region/country. It is updated after an ad is first served, and displays election ads from verified advertisers that have one or more impressions. The data from the AU Political Advertising Transparency Report and Ad Library is also available on [Google Cloud's BigQuery](#). Using BigQuery's API, any interested third party can write code and run their own unique queries on this data set to develop charts, graphs, tables, or other visualisations of election ads on Google platforms.

Within the Google Advertising Political Advertising Transparency team, a User Experience team has been devoted to understanding target users' (including researchers') needs and working with Product Managers, Engineers, and others to develop new and improve existing features and functionalities of the Political Advertising Transparency Report. For example, Google Advertising has an always on survey that surfaces for anyone who views the Political Advertising Transparency Report. Via this survey, Google Advertising consistently gets feedback from researchers, as well as other user groups, and uses this feedback to inform improvements to the website.

Case Study: Political Advertising Transparency Report insights

The [Political Advertising Transparency Report](#) can be filtered by country/region, date and ad format (text, image, video) to get insights into the top advertisers during specific election periods and their respective ad spend.

Only [verified advertisers](#) are permitted to run election ads in Australia. All election ads run by verified election advertisers must contain a [disclosure](#) that identifies who paid for the ad. During 2023:

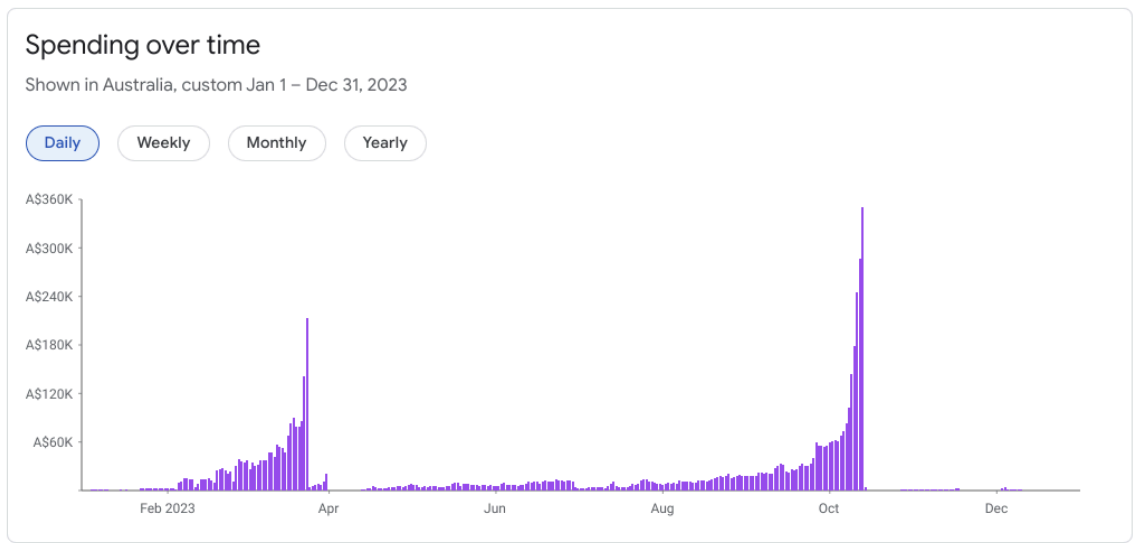
- **70 verified advertisers** ran Australia Election Ads in Australia
- **27,136 ads were rejected due to unverified advertisers** attempting to run Australia Election Ads.

The New South Wales state election was held on 25 March 2023, and the Referendum on an Aboriginal and Torres Strait Islander Voice was held on 14 October 2023. The below metrics summarise selected insights about Australia Election Ads shown in Australia in 2023.

- Total election ad spend amounted to **A\$6,016,800 (AUD)**, spent across **2,324 ads**
- The top three advertisers by ad spend were the Australians for Indigenous Constitutional Recognition Ltd (**A\$2.1M**), Australian Labor Party (New South Wales Branch) (**A\$1.3M**), and

the Australian Electoral Commission (**A\$728K**)

- The top three states or territories by ad spend were New South Wales (**A\$2.8M**), Victoria (**A\$1.0M**), and Queensland (**A\$847K**)
- Video ad formats accounted for the highest ad spend at **A\$3.9M**, followed by text ad formats (**A\$1.8M**) and image ad formats (**A\$263K**).



Source: [Political Advertising Transparency Report](#) as of March 28, 2024

Note: Google is continuously reviewing its data and improving its classification of election ads. As Google constantly increases the accuracy of our data, historical data may slightly fluctuate over time.

Objective 6: Strengthen public understanding of disinformation and misinformation through support of strategic research.

Outcome 6: Signatories support the efforts of independent researchers to improve public understanding of disinformation and misinformation.

We continue to support global and Australia-specific efforts to enhance awareness and detection of misinformation and promote authoritative sources of information. Examples of these collaborations are provided in the table below.

Name of university, institute, or company	Overview of research
Australian Associated Press (AAP)	AAP worked with Indigenous journalists to create fact-checks regarding the 2023 Referendum on an Aboriginal and Torres Strait Islander Voice. They then provided these fact-checks to Indigenous publications, free of charge. For more information, see here .
First Draft	First Draft worked with 10 news publishers (a mix of small and large) and additional community organisations to help them identify false, misleading and confusing claims during the 2022 Federal election campaign. They supported journalists through simulations and masterclasses and created an online community for the partner publishers. First Draft also provided daily alerts and weekly briefings on false and misleading claims that were circulating online. The partners they worked with included: Asian Alliance Australia, AAP, Canberra Times, SBS, NITV, Launceston Examiner, Northern Daily Leader Tamworth Wimmera Mail-Times, Codebreakers, Democracy in Colour, Crikey, Joy, Guardian Australia, 3CR Community Radio, RMIT ABC Fact Check, The Drum, ABC and The Humanism Project.
Squiz Kids	Google funds the development of Squiz Kids' media literacy program Newshounds which is delivered in Australian classrooms to help primary school children understand which media stories they should trust. Over 2,000 Australian classrooms are using the 8 part training module to

	<p>teach children how to ‘stop, think and check’ when they read or hear a story. The program has had early success, 86% of pilot students said Newsounds has changed the way they consume media and every teacher said they can see improvements in students’ abilities to critically consume media.</p>
YouGov & Poynter Report	<p>In 2022, Google Search partnered with YouGov and Poynter on a report (‘A Global Study on Information Literacy’) that summarised findings from a survey of 8,585 respondents from 7 countries around the world. The report focused on consumer habits and practices related to misinformation, search literacy, and information journeys - see the infographic and full findings.</p>
YouTube Researcher Program	<p>The YouTube Researcher Program provides scaled, expanded access to global video metadata across the entire public YouTube corpus via a Data API for eligible academic researchers from around the world, who are affiliated with an accredited, higher-learning institution. The program allows researchers to independently analyse the data they collect, including generating new/derived metrics for their research. Information available via the Data API includes video and channel title and description, views, likes, comments, channel metadata, search results, and other data.</p> <p>There is an application process with detailed policies, eligibility criteria, and guidance that can be found on the YouTube Research Policies page.</p>
Google Researcher Program	<p>As of 28 August 2023, eligible EU researchers can apply for access to publicly available data across some of Google’s products, including Search and YouTube, through the Google Researcher Program. Search and YouTube will provide eligible researchers (including non-academics that meet predefined eligibility criteria) with access to limited metadata scraping for public data. This program aims to enhance the public’s understanding of Google’s services and their impact. For additional details, see the Researcher Program landing page.</p>
Google Trends	<p>Google Search and YouTube provide publicly available data via Google Trends, which provides access to a largely unfiltered sample of actual search requests made to Google Search and YouTube’s search function. By sampling data, Google can look at a dataset representative of all Google and YouTube searches, while finding insights that can be processed within minutes of an event happening in the real world. It is anonymised (no one is personally identified), categorised (determined by the topic for a search query) and aggregated (grouped together). See Trends Help Centre for details.</p>

Objective 7: Signatories will publicise the measures they take to combat disinformation and misinformation.

Outcome 7: The public can access information about the measures Signatories have taken to combat disinformation and misinformation.

In [Objective 1 / Outcome 1d](#) of this report, we provide details (and links to corresponding materials) regarding publicly available and accessible information outlining measures we take to combat misinformation.

Additionally, Google published '[How we're fighting misinformation across Australia](#)' in March 2023 with additional details on how it is making it easier for people in Australia to evaluate information online.

Google has also made investments in developing and publishing videos to help advertisers and publishers understand our policies, including misinformation policies. These videos can be viewed on the [Google Transparency Center](#).

Concluding remarks

Through the framework of the Objectives and Outcomes set forth in the Australian Code of Practice on disinformation and misinformation, we have discussed how our products and services aim to mitigate the risk of harms arising from misinformation. We have outlined the various policies, safeguards and measures implemented across our products and services during 2023 to reduce the propagation and impact of misinformation. The case studies, examples and quantitative data points provided highlight the impact of these approaches. Additionally, we have detailed how we engage with the public and users, how we support other organisations tackling this issue, and how we provide further transparency regarding our own ongoing efforts.

We recognise that misinformation, and the risks associated with this issue, will continue to evolve. As such, we will continue to evaluate and adapt the measures and policies that we put in place across our products and services and invest in developing effective ways to protect our users and the integrity of our services. As we continue to report on the progress of this work, we look forward to engaging further with the Government, DIGI, the industry, civil society, users, academia and other key stakeholders on this issue.

Appendices

Appendix A: Google Misinformation/Disinformation Policies

Google Search & Google News Policies

- **[Google Search Webmaster Guidelines](#)**: Our webmaster guidelines prohibit techniques which could be used to deceive our ranking systems or abuse our users.
 - [Automatically generated content](#)
 - [Participation in link schemes](#)
 - [Cloaking](#)
 - [Sneaky redirects](#)
 - [Hidden text or links](#)
 - [Doorway pages](#)
 - [Scraped content](#)
 - [Loading pages with irrelevant keywords](#)
 - [Creating pages with malicious behaviour](#)
 - [Abusing structured data markup](#)
 - [Sending automated queries to Google](#)
- **[Google Search Content Policies](#)**: These policies apply to content surfaced anywhere within Google Search, which includes web results. These policies include additional Search Features policies that apply to Google News and other specialised features, which include but are not limited to:
 - **Deceptive Practices Policy**: This policy prohibits sites or accounts that impersonate any person or organisation, or that misrepresent or conceal their ownership or primary purpose. We do not allow sites or accounts that engage in inauthentic or coordinated behaviour that misleads users.
 - **Manipulated Media Policy**: This policy prohibits audio, video, or image content that's been manipulated to deceive, defraud, or mislead by means of creating a representation of actions or events that verifiably didn't take place.
 - **Medical Content Policy**: This policy does not allow content that contradicts or runs contrary to scientific or medical consensus and evidence-based best practices.
- **[Google News Policies](#)**: These content and behaviour policies help ensure a positive experience for users and publishing partners. Along with Google Search's overall Content Policies, Google News has additional feature-specific policies as noted below:
 - **Ads & Sponsored Content Policy**: Ads and other paid promotional material should not exceed content on pages. This policy states that we do not allow content that conceals or misrepresents sponsored content as independent, editorial content.

- **Misleading Content Policy:** This policy states that News does not allow preview content that misleads users to engage with it by promising details which are not reflected in the underlying content.
- **Transparency Policy:** This Google News Policy notes that news sources on Google should provide clear dates and bylines, as well as information about authors, the publication, the publisher, company or network behind it, and contact information.

Google Ads Policies

- **[Google Ads Policies](#):** These policies are designed not only to abide by laws but also to ensure a safe and positive experience for our users. This means that our policies prohibit some content that we believe to be harmful to users and the overall advertising ecosystem. Our policies cover four broad areas:
 - Prohibited content
 - Prohibited practices
 - Restricted content and features
 - Editorial and technical.
- **[Google Ads Misrepresentation Policy](#):** This policy prohibits content and behaviours that deceive users by excluding relevant product information or providing misleading information about products, services, or businesses. Violations of this policy, although not a comprehensive list, include unacceptable business practices, coordinated deceptive practices, unreliable claims (such as making claims that are demonstrably false and could significantly undermine participation or trust in an electoral or democratic process) and misleading representation.
 - **Unacceptable Business Practices Policy** does not allow:
 - Scamming users by concealing or misstating information about the advertiser's business, product, or service;
 - Ad destinations that use 'phishing' techniques to gather user information.
 - **Coordinated Deceptive Practices Policy** prohibits two practices:
 - Coordinating with other sites or accounts and concealing or misrepresenting your identity or other material details about yourself, where your content relates to politics, social issues, or matters of public concern;
 - Directing content about politics, social issues, or matters of public concern to users in a country other than your own, if you misrepresent or conceal your country of origin, or other material details about yourself.
 - **Misleading Representation Policy** prohibits advertisers from:
 - Making misleading statements, obscuring, or omitting material information about your identity, affiliations or qualifications;
 - Providing an inaccurate business name or business name that does not clearly represent the advertised business or disambiguates from similar businesses in the ad or user interactions.
 - **Manipulated Media Policy** does not allow advertisers to manipulate media to deceive, defraud, or mislead others.

- **Unreliable Claims Policy** does not allow advertisers to make inaccurate claims or claims that entice the user with an improbable result as the likely outcome a user can expect. This includes claims related to health and weight loss, financial products or money making schemes, and politics, social issues, or matters of public concern.
- **[Google Ads Inappropriate Content Policy](#)**: As noted in our help centre, this policy prohibits ads or destinations that display shocking content or promote hatred, intolerance, discrimination, or violence.

Google AdSense Policies

- **[Google Publisher Policies](#)**: Users who monetise content with Google must adhere to all Google Publisher Policies. Prohibited content and behaviours include but are not limited to: dangerous or derogatory content, misrepresentative content (such as making claims that are demonstrably false and could significantly undermine participation or trust in an electoral or democratic process) or coordinated deceptive practices.
- **[AdSense Program Policies](#)**: AdSense publishers are required to adhere to the AdSense Program Policies or risk their ad or account being disabled. It is the publisher's responsibility to keep up to date with and adhere to the following policies: invalid clicks and impressions, encouraging clicks or views (non-rewarded inventory), traffic sources, ad behaviour, ad placement, site behaviour, deceptive site navigation.

YouTube Policies

- **[YouTube Community Guidelines](#)**: These Community Guidelines outline what type of content is and is not allowed on YouTube. These policies apply to all types of content on our platform, including videos, comments, and links. These policies cover: spam & deceptive practices, violent or dangerous content, regulated goods, [harmful conspiracy theories](#), and misinformation among other areas.
- YouTube Misinformation Policies detailed below can be found on this [landing page](#).
 - **[YouTube Misinformation Policies](#)**: Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes certain types of misinformation that can cause real-world harm, like promoting harmful remedies or treatments, certain types of technically manipulated content, or content interfering with democratic processes.
 - **[YouTube Elections Misinformation Policies](#)**: Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes misinformation that can cause real-world harm, like certain types of technically manipulated content, and content interfering with free and fair democratic election processes.
 - **[YouTube Medical Misinformation Policy](#)**: YouTube does not allow content that poses a serious risk of egregious harm by spreading medical misinformation that contradicts

local health authorities' (LHAs) or the World Health Organisation's (WHO) guidance about specific health conditions and substances in relation to prevention, treatment or denial.

- **[YouTube Impersonation Policy](#)**: This policy states that content intended to impersonate a person or channel is not allowed on YouTube. YouTube also enforces trademark holder rights. When a channel, or content in the channel, causes confusion about the source of goods and services advertised, it may not be allowed.
- **[YouTube Fake Engagement Policy](#)**: YouTube does not allow anything that artificially increases the number of views, likes, comments, or other metrics either by using automatic systems or serving up videos to unsuspecting viewers. Content and channels that do not follow this policy may be terminated and removed from YouTube.
- **[YouTube Spam, Deceptive Practices, and Scam Policies](#)**: YouTube does not allow spam, scams, or other deceptive practices that take advantage of the YouTube community. We also do not allow content where the main purpose is to trick users into leaving YouTube for another site.
- **[YouTube Advertiser-friendly Content Guidelines](#)**: Users in the YouTube Partner Program can share revenue from ads. This policy exemplifies content that is not suitable for ads that will result in a 'limited or no ads' monetisation state.
- **[YouTube Channel Monetisation Policies](#)**: YouTube Monetisation Policies include YouTube's Community Guidelines, Terms of Service, Copyright, and Google AdSense Program Policies. YouTube enforces these Monetisation Policies by:
 - Turning off Ads from your content;
 - Suspending your participation in the YouTube Partner Program; and
 - Suspending or even terminating your YouTube channel.

Appendix B: Tools and features used to empower users in making informed choices

Google has developed many features and interventions to provide more context to users and ensure that authoritative sources are elevated in response to searches or browsing intents related to health, civic participation, current events, and other topics where users want content that they can trust. The features and interventions listed below are all available to Australian users of our services.

- **Surfacing Fact-Checks on Google Search, Images, and News**: easily enable users to find articles that fact-check public information;
- **Elevating original reporting in Google Search**: provide greater context to subsequent news stories;

- **'Full Coverage' in Google News:** help users access context and diverse perspectives about the news stories they read;
- **'Information Panels', including those providing topical context on YouTube:** provide greater context around topics related to searches or videos (additional information can be found [here](#));
- **'Breaking News' and 'Top News Shelves' on YouTube:** prominently surface news content from authoritative sources only;
- **Authoritativeness in YouTube recommendations:** prioritise information from authoritative sources for newsworthy events and topics prone to misinformation in search results and recommendations (additional information can be found [here](#)); and
- **Labelling state-funded news channels on YouTube:** label publishers that are government or publicly funded on the watch pages of their videos.

Appendix C: Advertiser Verification Process

Google's advertiser verification program verifies advertiser identity, then discloses the advertiser behind specific ads in the ['Why this Ad' \(in some places, 'About this Ad'\)](#) menu. This verification and disclosure feature applies to the ads that Google shows on our own properties and those of our publishing partners. Note that the disclosures view can vary slightly between products. Google users can control what types of ads they see, and whether they would like to see more or less of any specific advertiser, through [My Ad Centre](#).

Verification process for Australian election advertisers

Prospective advertisers who wish to run election ads in Australia are also required to go through a verification process. Once Google verifies the advertiser's eligibility to run election ads, they receive an email and an in-account notification. Verifying their identity may require two steps and each step can take up to 5 business days. Our teams are trained to handle this process at scale across Australia, and are equipped to respond to related questions from the political parties and candidates participating in, and institutions responsible for, Australian elections.

Meta response to the Australian Code of Practice on Disinformation and Misinformation

Reporting period: January - December 2023

Summary

As a founding member and signatory to the Australian Code of Practice on Disinformation and Misinformation (**the Code**), Meta provides our latest transparency report to publicise the measures that we take to combat disinformation and misinformation.

This latest update outlines Meta's approach to combat disinformation and misinformation in Australia during the 2023 calendar year. There are several key trends and changes that have occurred during the reporting period:

- **Global adversarial threats:** In 2023, we disrupted a number of Co-ordinated Inauthentic Behaviour (CIB) networks. More than half of these CIB networks targeted audiences outside of their countries of operation. We removed the majority of these networks before they were able to build authentic audiences. We identified four trends in CIB throughout 2023, specifically: an increase in China-based CIB disruptions, for-hire surveillance operations being behind CIB globally, abuse of domain name infrastructure and the Russian network 'Doppelgänger' trying to stay online.
- **Changed public health emergency declarations:** As the COVID-19 situation evolved, in mid-2022, we sought advice from Meta's Oversight Board specifically on whether we should change our approach to COVID-19-related misinformation. The Oversight Board issued an opinion finding that Meta should prepare measures for when the World Health Organisation lifted its public health emergency declaration, in order to protect freedom of expression and other human rights. In June 2023, we released our response to the Oversight Board's recommendations,¹ announcing that we would take a more tailored approach to our COVID-19 misinformation rules, in line with the Board's recommendations: in countries that have a COVID-19 public health emergency declaration, we will continue to remove content for violating our COVID-19 misinformation policies, given the risk of imminent physical harm. Globally, our COVID-19 misinformation rules are no longer in effect, as the global public health emergency declaration that triggered those rules has been lifted. Our commitments under the Code have therefore been

¹ Oversight Board, , ['Oversight Board publishes policy advisory opinion on the removal of COVID-19 misinformation'](#), April 2023.

adjusted to reflect this policy change.

- **Broader deployment of Generative AI:** Throughout 2023 and ongoing in 2024, there has been an increased awareness of the use of AI technology as part of content governance systems and also an increased focus on distribution, and use by businesses and consumers, of Generative AI technology. To respond to questions about responsible innovation, we began re-evaluating our policies and consulted with a range of stakeholders globally to identify if we needed to update and amend these to keep pace with the rapid generative AI advances. These consultations, as well as new recommendations by Meta's Oversight Board, led to a range of new changes specifically relating to how we handle manipulated media.

Looking ahead to 2024, Meta has opted into 40 commitments under the code and since the Code took effect in 2021, we have released three transparency reports outlining our efforts to meet these commitments.

In 2023, Meta's transparency report outlined 38 commitments to meet our obligations across both Facebook and Instagram. This report outlines the steps we took to meet these commitments across the reporting period of January to December 2023.² We have provided new case studies on the global trends in Coordinated Inauthentic Behaviour for 2023, as well as our work during the 2023 Aboriginal and Torres Strait Islander Voice Constitutional Referendum, where we introduced a number of specific measures to combat misinformation and disinformation.

We also continue to release up-to-date, Australia-specific data on our misinformation efforts to maintain transparency with the community and policymakers about the impact of these misinformation efforts. Between 1 January and 31 December 2023:

- We took action on over 9,700 pieces of content across Facebook and Instagram in Australia for violating our Misinformation policies.
- In addition to this, over 6,200 ads were removed in Australia for violating our Misinformation policy.
- We displayed warnings on over 9.2 million distinct pieces of content on Facebook, and over 510,000 on Instagram, in Australia (including reshares) based on articles written by our third-party fact-checking partners.
- We removed over 75,000 ads in Australia for not complying with our Social Issues, Elections and Politics (SIEP) ads policy.

In 2023, some further highlights of our work included:

- Implementing a suite of measures in the lead up to the 2023 Voice Referendum to proactively detect and remove content that breaches our policies, combat

² This reporting period has been specified in the guidelines provided by the independent reviewer, engaged by DIGI.

misinformation, harmful content and coordinated inauthentic behaviour, and promote civic participation.

An example of these efforts are the referendum day reminders we deployed to remind people to vote in the referendum.

- Our Civic Products reached around 12.7 million users on Facebook with around 13.6 million impressions.
- For Instagram, around 6.55 million users were reached with around 6.57 million impressions.
- Investing in new research and analysis relating to information integrity surrounding the referendum debate and its outcomes, including in partnership with La Trobe University and RMIT CrossCheck.
- Sponsoring and hosting the Australian Media Literacy Alliance's Australian Media Literacy Summit, which brought together a range of journalists, academics, educators, librarians and other experts to discuss and learn about diverse ways to strengthen media literacy education in Australia.
- Supporting Western Sydney University's 2023 Young People and News longitudinal survey, which provides findings about the news attitudes, practices and experiences of young Australians aged between 8-16 years.
- Introducing new tools to provide users with more transparency around content ranking algorithms and recommendation systems, including updates to the Facebook and Instagram System Cards.
- We conducted significant research and collaborated with key industry partners across 2023 to produce a set of signals that will help us to label content that we believe has been made with AI. We believe providing transparency and additional context is now the better way to address manipulated media and avoid the risk of unnecessarily restricting freedom of speech, so we'll keep this content on our platforms so we can add labels and context.
- Preparing for and announcing a new AI Disclosure policy that allows advertisers to label Social Issues, Elections, and Politics (SIEP) Ads when they have been digitally created or altered so that viewers are aware. This policy went into effect in 2024.
- Sponsoring events in relation to promoting authoritative information, including a new training series called 'Connect, Alert, Inform', for emergency response organisations across Australia, New Zealand and the Pacific Islands. The training focused on helping to strengthen emergency communicators' skills in using Meta's services to build community and deliver critical disaster-related information. It included a module specifically developed and delivered by RMIT CrossCheck for

disaster communicators on combating, addressing and avoiding the amplification of misinformation surrounding a disaster or emergency situation.

In support of the Code and making meaningful developments in our transparency efforts, we have included two new metrics to this report. These are:

- The number of ads removed for violating our Misinformation policy from January to December 2023 in Australia.
- The number of ads removed for not complying with Meta’s SIEP advertising policies,³ from January to December 2023 in Australia.

These steps are part of our extensive global efforts to combat misinformation and disinformation.

This report also outlines Meta’s 40 commitments for the next reporting period. Meta will continue to document and report on relevant updates, changes and developments to our integrity practices related to disinformation and misinformation.

We look forward to continuing to work with Australian policymakers, civil society, academics and experts on steps to combat misinformation and disinformation in Australia over the next year.

³ Meta, [‘Ads about Social Issues, Elections or Politics’](#), Meta Transparency Centre, 19 March 2024

List of Meta commitments under the Australian industry code on misinformation and disinformation for 2024

A copy of the 2023 commitments can be found in **Appendix A**.

The 40 commitments for 2024 are outlined below:

<p>Outcome 1: Combatting misinformation / disinformation</p>	<p>Outcome 1a. Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures.</p> <ul style="list-style-type: none">• Meta removes networks of accounts, Pages and Groups that violate our inauthentic behaviour policy, including disinformation that violates our policy on Inauthentic Behaviour.• Meta provides transparency about accounts, Pages and Groups removed under our Inauthentic Behaviour policy.• Meta partners with experts and organisations who assist in providing tips or further investigation about possible inauthentic behaviour on our services.• Meta removes misinformation that violates the misinformation section of our Community Standards.• Meta removes manipulated videos, also known as “deepfakes”, that violate our Manipulated Media policy.• Meta removes election-related misinformation that may constitute voter fraud or interference under our Misinformation, Coordinating Harm and Promoting Crime policies.• Meta removes election-related misinformation that may constitute voter suppression.• Meta removes fake accounts.• Meta allows for appeals in instances where users may disagree with our enforcement, including to the independent and external Oversight Board.• Meta partners with third-party fact-checking organisations, globally and in Australia, to assess the accuracy of content on our services.• Meta applies a warning label to content found to be false, partly false, altered, or missing context by third-party fact-checking organisations.
--	---

- Meta reduces the distribution of content found to be false, partly false, or altered, by third-party fact-checking organisations.
- Meta proactively searches for content that matches content debunked by our fact-checking partners, to apply the same treatments.
- Meta takes action on Pages, Groups, accounts or websites found to repeatedly share misinformation.
- Meta continues to build safeguards into our GenAI features and models so that people can have a safer and more enjoyable experience.

Outcome 1b. Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.

- Meta makes information available via a dedicated website that outlines our efforts to combat misinformation.

Outcome 1c. Users can report content and behaviours to Signatories that violate their policies under 5.10 through publicly available and accessible reporting tools.

- Meta makes on-platform reporting channels available to users for false information.
- Meta's Australian third-party fact-checking partners are also able to receive referrals from the public.

Outcome 1d. Users will be able to access general information about Signatories' actions in response to reports made under 5.11.

- Meta makes global transparency reports available regularly.
- Meta will supplement these reports with additional Australia-specific statistics, provided as part of this Annual Report process.

Outcome 1e. Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.

	<ul style="list-style-type: none"> • Meta will continue to provide ongoing transparency of our content ranking algorithms and give users more control over the content they see. • Meta takes steps to limit the possible distribution of misinformation via recommendations.
<p>Outcome 2: Disrupt monetisation and advertising incentives</p>	<ul style="list-style-type: none"> • Meta sets a higher threshold for users to be able to advertise on our services, and takes action against users who spread misinformation. • Meta commits to defund the dissemination of misinformation.
<p>Outcome 3: Combat inauthentic user behaviour</p>	<p>See items listed under Outcome 1.</p>
<p>Outcome 4: Empower consumers to be informed</p>	<ul style="list-style-type: none"> • Meta provides contextual information around posts that users see from public Pages. • Meta provides consumers with on-platform tools to understand why they are seeing a particular post and to manage the content they see on Facebook and Instagram. • Meta reduces the distribution and recommendation of problematic and low quality content on Facebook and Instagram. • Meta will provide ongoing transparency in relation to AI-generated images and content. • Meta will look for opportunities to work with the Australian Government or local organisations to promote authoritative information and/or media literacy.
<p>Outcome 5: Political advertising</p>	<ul style="list-style-type: none"> • Meta requires all advertisers of political and social issue ads to complete an ad authorisation, which includes verifying the advertiser’s identity. • Meta requires political and social issue ads to include a disclaimer disclosing who is paying for the ad. • Meta provides the Ad Library, a searchable archive of all political and social issue ads on our services in Australia.

	<ul style="list-style-type: none"> • Meta enables an Ad Library report that provides aggregated spend information about Pages undertaking political and social issue ads. • Meta will explore ways to ensure that there is appropriate ads transparency in relation to AI.
<p>Outcome 6: Research</p>	<ul style="list-style-type: none"> • Meta supports research and events in relation to misinformation and media literacy. • Meta supports research and events in relation to disinformation. • Meta collaborates with researchers to undertake surveys of our users to assess their views on relevant social issues. • Meta will make the Meta Content Library and API research tools available to third-party fact-checking partners and qualified users. • Meta will continue to support new areas of research relating to misinformation and disinformation in 2024.
<p>Outcome 7: Annual reports</p>	<ul style="list-style-type: none"> • Meta will continue to publish annual reports in Australia, such as these, to be transparent about the steps we are taking to combat disinformation and misinformation.

Table of contents

Summary	1
List of Meta commitments under the Australian industry code on misinformation and disinformation for 2024	5
Table of contents	9
Reporting against 2023 commitments	10
Outcome 1a	10
Case study: Meta’s response to global CIB attempts and global adversarial threats throughout 2023.	11
Outcome 1b	20
Outcome 1c	21
Outcome 1d	22
Outcome 1e	24
Outcome 2	26
Outcome 3	27
Outcome 4	28
Case study: Aboriginal and Torres Strait Islander Voice referendum	30
Outcome 5	35
Outcome 6	37
Outcome 7	40
Appendix A: 2023 specific commitments made by Meta under the industry code on misinformation and disinformation	41

Reporting against 2023 commitments

Outcome 1a

Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures.

Signatories will develop and implement measures which aim to reduce the propagation of and potential exposure of users of their services and products to Disinformation and Misinformation

Our approach to misinformation and disinformation is consistent with that which was outlined in our 2021, 2022 and 2023 transparency reports.⁴ Below we outline developments from 2023:

- **Meta removes networks of accounts, Pages and Groups that violate our inauthentic behaviour policy, including disinformation that violates our policy on Inauthentic Behaviour.**

We have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports. As mentioned in the baseline 2021 report, Meta's approach focuses on removing misinformation which could directly contribute to the risk of imminent physical harm, misinformation that could interfere with the functioning of political processes, certain highly deceptive manipulated media, as well as fake accounts (which are often vehicles for disinformation). It also focuses on reducing the spread of other misinformation, while promoting authoritative information. Further details of Meta's most recent efforts to remove coordinated inauthentic behaviour (CIB) are outlined in the commitment below.

- **Meta provides transparency about accounts, Pages and Groups removed under our Inauthentic Behaviour policy.**

We use a combination of policies, tools, expert teams and partnerships to detect and remove networks of inauthentic behaviour (IB) and CIB - both foreign and domestic.

We continue to report on our efforts to disrupt CIB through our Community Standards Enforcement Report and Quarterly Adversarial Threats report.⁵ As mentioned in our most recent report, Meta has removed over 200 covert influence

⁴ Meta, '[Meta's response to the Australian disinformation and misinformation industry code](#)', *Meta Australia blog*, May 2021.

⁵ Meta, [Community Standards Enforcement Report](#), and Meta, [Adversarial Threats report](#), Meta Transparency Centre.

operations between 2017-2022.⁶

In our 2023 transparency report, we reported that Meta had taken action on four instances of CIB operations that targeted Australians, since 2017. This number increased in 2023, to five instances as a China based network targeted Australia and other western countries. During the 2023 Aboriginal and Torres Strait Islander Voice Referendum, we did not see any evidence of coordinated inauthentic behaviour targeting Australia.

Below we provide a case study that looks at the adversarial threat trends globally across 2023.

Case study: Meta's response to global CIB attempts and global adversarial threats throughout 2023

A high level summary of CIB Networks disrupted in 2023:

- In Q1 we removed six separate covert influence operations for violating our policy against CIB. They originated in:
 - The United States & Venezuela
 - Iran
 - China (two separate networks originated in China)
 - Georgia
 - Burkina Faso & Togo
- In Q2 we removed three separate covert influence operations that violated our policy against CIB. They originated in:
 - Türkiye
 - Iran
 - China
- In Q3 we removed three separate covert influence operations that violated our policy against CIB. They originated in:
 - China (two separate networks originated in China)
 - Russia
- In Q4 we removed three separate covert influence operations that violated our policy against CIB. They originated in:
 - China
 - Myanmar
 - Ukraine

⁶ Meta, D Agranovich, '[Recapping our 2022 coordinated inauthentic behaviour enforcements](#)', *Meta Newsroom*, 15 December 2022.

More than half of these CIB networks targeted audiences outside of their countries of operation. We removed the majority of these networks before they were able to build authentic audiences.

Trends in coordinated inauthentic behaviour throughout 2023:

- **An increase in China based CIB disruptions:**
 - In Q2 2023, we removed a network of 7,704 Facebook accounts, 954 Pages, 15 Groups and 15 Instagram accounts for violating our policy against coordinated inauthentic behaviour. This network originated in China and targeted many regions around the world, including Australia, the United Kingdom, Taiwan, the United States, Japan, and global Chinese-speaking audiences.
 - We began this investigation after reviewing public reporting about off-platform activity that targeted a human rights NGO in late 2022. Following this lead, we were able to uncover a large and prolific covert influence operation which was active on more than 50 platforms and forums, including X (formerly Twitter), YouTube, TikTok, Reddit, Pinterest, Medium, Blogspot, LiveJournal, VKontakte, Vimeo, and dozens of smaller platforms and forums, as well as Facebook and Instagram.
 - On our platforms, this network was run by geographically dispersed operators across China who appear to have been centrally provisioned with internet access and content directions. Many of their accounts were detected and disabled by our automated systems. We assess that this likely led the people behind it to increasingly shift to posting its content on smaller platforms and then trying to amplify it on larger services in hopes to maintain persistence.
 - We have not found evidence of this network getting any substantial engagement among authentic communities on our services. This network typically posted positive commentary about China and its province Xinjiang and criticisms of the United States, Western foreign policies, and critics of the Chinese government including journalists and researchers.
 - Other China based networks identified in 2023 targeted India and Tibet, the United States, the EU, Taiwan, Sub-Saharan Africa, Japan, Central Asia and the Uyghur community around the world.

- Networks have utilised content in English and Chinese.
- **For-hire operations:**
 - Our 2023 reporting called out a trend that we continue to see, which is, for-hire surveillance organisations being behind covert influence operations globally, with many of the operations covered in our 4 quarterly reports being attributed to private entities. This included examples such as an IT company in China, a marketing firm in the United States and a political marketing consultancy in the Central African Republic, as well as several mercenary spyware companies operating in the EU despite EU privacy standards and regulations.⁷
- **Countering domain name abuse globally:**
 - Many threat actors continue to utilise domain name infrastructure in their malicious operations across the internet – from cyber espionage to covert influence campaigns and spyware firms. In 2023, we resolved a legal case against Freenom, a country code domain registry provider, whose domain names accounted for over half of all phishing attacks involving country code top-level domains (ccTLDs). This settlement resulted in Freenom announcing that it will exit the domain name business, including its operation of the country-code registries. While Freenom winds down its domain name business, it has agreed to treat Meta as a trusted notifier and it will also implement a block list to address future phishing, DNS abuse, and cybersquatting.⁸
- **The Russian network ‘Doppelganger’s’ persistence in trying to stay online:**
 - In 2022, we shared our threat research into the CIB network focused on supporting Russia’s invasion of Ukraine, dubbed ‘Doppelganger’, that operated across the internet, including running a large network of websites spoofing legitimate news outlets. In December 2022, we attributed it to two companies in Russia: Structura National Technology and Social Design Agency (Агентство Социального Проектирования). We banned these firms from our services. They were also later sanctioned by the EU.⁹ Our analysis from 2023 noticed several trends:

⁷ Meta, [Adversarial Threats report Q1 2023 and Q4 2023](#), Meta transparency Centre.

⁸ Meta, [Adversarial Threats report Q4 2023](#), Meta transparency Centre.

⁹ Meta, [Adversarial Threats report Q2 2023](#), Meta transparency Centre.

- Persistence: We assessed this campaign to be the largest and most aggressively persistent covert influence operation from Russia that has been seen since 2017. Since our initial disruption and continuous scrutiny by platforms and researchers, Doppelganger continued to create new domains in an attempt to evade detection. Given the nature of this operation and the type of entities behind it, this is expected behaviour across our industry with any CIB network we each take down.
- In addition to ongoing detection by our automated systems, our team has been monitoring and taking action against these recidivist attempts, and sharing findings with our peers and with the public. In total (as of Q3 2023), we've blocked over 2,000 of the operation's domains from being shared on our platform. We also blocked tens of thousands of attempts to run fake accounts and Pages on Meta's platforms.
- High input – low output: This operation stands out for the sheer wastefulness of its large-scale efforts. We expect Doppelganger to keep at it with its “smash-and-grab” approach by throwing a large amount of resources – even if it leads to a very high detection rate and loss of assets, as we described in September 2022 and in our last report.
- Expanding targeting, yet single mission: With Doppelganger focusing on weakening support for Ukraine against the Russia invasion, this operation appears to be trying to pick off some of Ukraine's international allies over time. Judging by the origin of the organisations that this operation spoofed, among other factors, this Russian campaign has expanded beyond targeting France, Germany and Ukraine itself for the first 8+ months to include the US and Israel earlier this year. While the exact reasoning behind this expansion is unknown, it likely reflects the fluid tasking of this operation (by its clients) and its single-minded mission.

- **Meta partners with experts and organisations who assist in providing tips or further investigation about possible inauthentic behaviour on our services.**

Meta continues to maintain close relationships with experts and organisations around the world so we can share threat findings and adapt our enforcement.

- **Meta removes misinformation that violates the misinformation section of our Community Standards, and will review its Misinformation and Harm Policy, in line with recommendations from the Oversight Board.**

Between 1 January and 31 December 2023, we took action on over 9,700 pieces of content across Facebook and Instagram in Australia for violating our Misinformation policies. This is compared to 91,000 pieces of content for the 2022 reporting period and 180,000 (globally) in 2021.

The figure for 2023 above reflects that during this reporting period – as mentioned above – the Oversight Board provided a policy advisory opinion on the removal of COVID-19 misinformation, to which Meta subsequently responded. The Board prepared its advice in response to Meta’s request on whether measures to address dangerous COVID-19 misinformation, introduced in extraordinary circumstances at the onset of the pandemic, should remain in place as many, though not all, countries around the world were seeking to return to more normal life. As a result, in June 2023, Meta updated our Misinformation and Harm policy to reflect the Oversight Board’s guidance. We now only remove this content in countries with an active COVID-19 public health emergency declaration. This change has impacted our enforcement metrics on removals for this reporting period, but does not change our overall approach to fact-checking. In general also, changes to this metric are an expected part of fluctuating content trends online.

- **Meta removes manipulated videos, also known as “deepfakes”, that violate our Manipulated Media policy.**

Throughout 2023, Meta continued to enforce its policy on manipulated videos, as outlined in our 2021, 2022 and 2023 transparency reports.

However, in the first half of 2023, we began reevaluating our policies to see if we needed a new approach to keep pace with rapid advances in generative AI technologies and usage. We completed consultations with over 120 stakeholders in 34 countries in every major region of the world. Overall, we heard broad support for labelling AI-generated content and strong support for a more prominent label in high-risk scenarios. Many stakeholders were receptive to the concept of people self-disclosing content as AI-generated.¹⁰

A majority of stakeholders agreed that removal should be limited to only the highest risk scenarios where content can be tied to harm, since generative AI is becoming a mainstream tool for creative expression. This aligns with the principles behind our Community Standards – that people should be free to express themselves while also remaining safe on our services.

¹⁰ Meta, [‘Our approach to labelling AI generated content and manipulated media’](#), Meta Newsroom, April 5 2024.

We also conducted public opinion research with more than 23,000 respondents in 13 countries and asked people how social media companies, such as Meta, should approach AI-generated content on their platforms. A large majority (82%) favour warning labels for AI-generated content that depicts people saying things they did not say.¹¹

Additionally, the Oversight Board noted that their recommendations were informed by consultations with civil-society organisations, academics, inter-governmental organisations and other experts.

We conducted this research throughout 2023 and starting from May 2024, we will begin labelling a wider range of video, audio and image content as “Made with AI” when we detect industry standard AI image indicators or when people disclose that they’re uploading AI-generated content.¹²

- **Meta removes election-related misinformation that may constitute voter fraud or interference under our Coordinating Harm and Promoting Crime policy.**

We have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports.

We continue to maintain our approach to remove content that may cause voter suppression or misrepresentation of election related information as outlined in our 2021, 2022 and 2023 transparency reports.

- **Meta removes fake accounts.**

Meta continues to enforce its policy on fake accounts. We do not allow fake accounts on Facebook and Instagram, as they can be vehicles for a range of harmful content and behaviour.

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.

From January to December 2023, we detected and removed 2.6 billion fake accounts on Facebook, on average we proactively detected and removed 98.95% of these accounts before they were reported to us.¹³ These are often caught within minutes of registration.

¹¹ Meta, [‘Our approach to labelling AI generated content and manipulated media’](#), Meta Newsroom, April 5 2024.

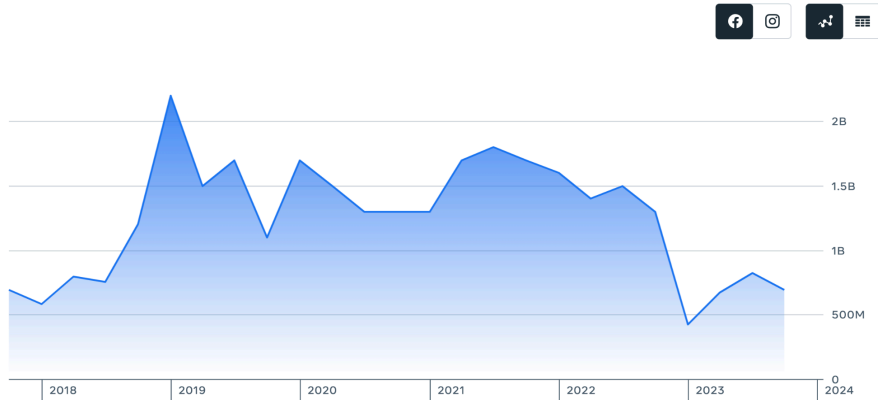
¹² Meta, [‘Our approach to labelling AI generated content and manipulated media’](#), Meta Newsroom, April 5 2024.

¹³ Meta, [‘Community Standards Enforcement Report’](#), Meta Transparency Centre.

of fake accounts we've taken action on (Oct 2017 - Dec 2023)¹⁴

ACCOUNTS ACTIONED

How many fake accounts did we take action on?



- **Meta allows for appeals in instances where users may disagree with our enforcement, including to the independent and external Oversight Board.**

As mentioned in the 2021, 2022 and 2023 transparency report, Meta has voluntarily established the independent Oversight Board to ensure greater accountability for our content governance on Facebook and Instagram.

In our last report, we noted that the Oversight Board issued 36 case decisions and 176 recommendations to Meta for future improvements.

Between January and December 2023, the Oversight Board issued 53 decisions and 66 recommendations.¹⁵

The Oversight Board also investigated two cases relating to the Australian Electoral Commission's (AEC) voting rules around the 2023 Aboriginal and Torres Strait Islander Voice Referendum.¹⁶ The cases involved two separate Facebook posts containing the same screenshot of information posted on X by the AEC ahead of the Referendum. The information shown included the message that: 'If someone votes at two different polling places within their electorate, and places their formal vote in the ballot box at each polling place, their vote is counted.' The posts were removed by Meta for violating the rule in our Coordinating Harm and Promoting Crime Community Standard that prohibits content calling for illegal

¹⁴ Note - graph represents number of fake accounts removed globally. Australia-specific statistics are not available.

¹⁵ Oversight Board, '[H2 2023 Transparency Report](#)', and, Oversight Board, '[Q4 2023 Transparency report](#)'.

¹⁶ Oversight Board, '[Oversight Board announces cases involving the Australian Electoral Commission's voting rules](#)', Oversight Board News, February 2024.

participation in a voting process. Both users had appealed the removal of these posts.

On 9 May 2024, the Oversight Board published its decision on the two cases, in which it upheld Meta's decision to remove the posts, as the two users' 'calls for others to engage in illegal behaviour impacted the political rights of people living in Australia' and that Meta 'was correct to protect democratic processes by preventing voter fraud attempts from circulating on its platforms, given the frequent claims that the Voice Referendum was rigged'.¹⁷

- **Meta partners with third-party fact-checking organisations, globally and in Australia, to assess the accuracy of content on our services.**

Meta has built a large global fact-checking network. As of December 2023, Meta partners with over 90 fact-checking partners covering more than 60 languages globally.¹⁸

Australians benefit from Meta's international approach to fact-checking - an Australian user will see a warning label on content that has been fact-checked by a third-party fact-checking partner. Content found to be false by our international fact-checking partners (including our Australian fact-checkers listed below) will be demoted in an Australian user's Feed, and will not be recommended, meaning there is less chance of them seeing it.

In Australia, we partner with three third-party fact-checkers: Australian Associated Press (AAP), Agence France Presse (AFP) and RMIT FactLab.¹⁹ In preparation for the 2023 Aboriginal and Torres Strait Islander Voice Referendum, we provided a one-off funding boost to AAP and AFP, so that they could increase their capacity in the lead up to the referendum.²⁰

- **Meta applies a warning label to content found to be false by third-party fact-checking organisations.**

We have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports.

¹⁷ Oversight Board, '[Oversight Board upholds Meta's decisions in Australian Electoral Commission voting rules cases](#)', May 2024.

¹⁸ Meta, '[Where we have fact-checking](#)', *Meta for Media*.

¹⁹ *Note:* From 28 August 2023, RMIT Factlab was suspended from our third-party fact checking program pending the International Fact-Checking Network's decision on RMIT FactLab's expired certification. Following the IFCN's review and RMIT FactLab's recertification on 6 Nov 2023, Meta has since reinstated RMIT FactLab to its third-party fact-checking program.

²⁰ Meta, '[How Meta is preparing for the Voice to Parliament Referendum](#)', *Medium*, 9 July 2023.

The focus of Meta’s fact-checking program is to identify and address viral misinformation, particularly clear hoaxes that have no basis in fact. Fact-checking partners prioritise probably false claims that are timely, trending and consequential. Meta applies a warning label to content found to be misinformation by the fact-checking organisations. Once fact-checkers have determined that a piece of content contains misinformation, Meta uses technology to identify near-identical versions of that content across Facebook, Instagram and Threads. Fact-check labels are only applied to near-identical versions of content that has already been rated; this means that labels are not applied to content that makes a similar claim but is differently expressed.²¹

Between 1 January and 31 December 2023, we displayed warnings on over 9.2 million distinct pieces of content on Facebook, and over 510,000 on Instagram, in Australia (including reshares) based on articles written by our third-party fact-checking partners. This compares with 9 million on Facebook in Australia for the 2022 reporting period, and a global number for 2021 on Facebook which was 190 million.

- **Meta reduces the distribution of content found to be false by third-party fact-checking organisations.**

We have maintained the approach outlined in our 2021, 2022 and 2023 transparency report.

- **Meta proactively searches for content that matches content debunked by our fact-checking partners, to apply the same treatments.**

Meta uses our technology to detect content that is the same or near-identical versions to content that has been fact-checked by our fact-checking partners. Generally, we do not apply treatments such as warning labels to content that makes a similar claim rated by fact-checkers, if the content is not identical. This is because small differences in how a claim is phrased might change whether it is true or false.

Between January – December 2023, we displayed warnings on over 9.2 million distinct pieces of content on Facebook (including re-shares) and 510,000 on Instagram. This compares with 9 million on Facebook in Australia for the 2022

²¹ Following the submission and publication of Meta’s 2022 transparency report in May 2023, a complaint was submitted by Reset Australia to the Independent Complaints Sub-Committee (under the Australian Code of Practice on Disinformation and Misinformation). The Committee has dismissed the complaint on the grounds that Reset Australia produced no convincing evidence that Meta’s transparency report contained false statements (additional information on the complaint can be found [here](#)). To provide further clarity and transparency, Meta has updated the language in this report to ensure additional clarity regarding our third-party fact-checking program moving forward.

reporting period, and a global number for 2021 on Facebook which was 190 million.

- **Meta takes action on Pages, Groups, accounts, or websites found to repeatedly share misinformation.** We have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports.

Outcome 1b

Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.

Signatories will implement and publish policies and procedures and any appropriate guidelines or information relating to the prohibition and/or management of user behaviours that may propagate Disinformation and Misinformation via their services or products.

- **Meta makes available a detailed list of claims that we consider to violate our COVID-19 Misinformation & Harm Policy.**

As outlined in our previous transparency reports, during the pandemic, Meta made available a detailed list of claims that violated our COVID-19 Misinformation & Harm Policy. These claims were available in our help centre throughout 2023.²²

However, Meta's policies in relation to COVID-19 misinformation changed in June 2023 following advice and consultation from Meta's independent oversight board.²³ Our COVID-19 misinformation rules are no longer in effect globally as the global public health emergency declaration that triggered those rules has been lifted. During the pandemic, we removed content making one of 80 false claims about COVID-19 which public health authorities concluded were likely to directly contribute to the risk of imminent physical harm. We now only remove this content in countries with an active COVID-19 public health emergency declaration. This change has impacted our enforcement metrics for this policy globally.

- **Meta makes information available via a dedicated website that outlines our efforts to combat misinformation.**

²² Meta, '[COVID 19 and Vaccine Policy Updates and protections](#)', Facebook Help Centre.

²³ Oversight Board, '[Removal of COVID-19 Misinformation](#)', Oversight Board Decisions.

We have maintained the approach outlined in our 2021, 2022 and 2023 transparency report. Meta's policies on misinformation can be found in our Transparency Centre.²⁴

Outcome 1c

Users can report content and behaviours to Signatories that violate their policies under 5.10 through publicly available and accessible reporting tools.

Signatories will implement and publish policies, procedures and any appropriate guidelines or information regarding the reporting of the types of content and behaviours that may propagate Disinformation and Misinformation via their platforms.

- Meta makes on-platform reporting channels available to users for false information.

Meta continues to make reporting channels available to users on both Facebook and Instagram. As outlined in our 2021, 2022 and 2023 transparency reports, users can follow the following steps to report misinformation:

- Facebook Help Centre - <https://www.facebook.com/help/572838089565953>
- Instagram Help Centre - <https://www.facebook.com/help/instagram/2442045389198631>

Our Australian third-party fact-checking partners are also able to receive referrals from the public using the channels below:

- Agence France Presse: <https://factcheck.afp.com/contact>
- Australian Associated Press: <https://www.aap.com.au/make-a-submission/>
- RMIT FactLab: <https://www.rmit.edu.au/about/schools-colleges/media-and-communication/industry/factlab/debunking-misinformation>.

²⁴ Meta, '[Community Standards - Misinformation](#)', Meta Transparency Centre.

Outcome 1d

Users will be able to access general information about Signatories' actions in response to reports made under 5.11.

Signatories will implement and publish policies, procedures and/or aggregated reports (including summaries of reports made under 5.11) regarding the detection and removal of content that violates platform policies, including but not necessarily limited to content on their platforms that qualifies as Misinformation and/or Disinformation.

- **Meta makes global transparency reports available regularly.**

Meta continues to make aggregated reports publicly available on our misinformation and disinformation efforts to provide oversight of our work. A range of compliance, regulatory and proactive transparency reports can be found in our online Transparency Centre.²⁵

As outlined in our 2021, 2022 and 2023 transparency reports, each quarter, we report on metrics for preventing and taking action on content that goes against our Community Standards. This is outlined in Meta's Community Standards Enforcement Report (CSER).²⁶

In 2022, Meta published the results of an independent audit, conducted by Ernst & Young (EY), into Meta's CSER reporting. EY 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.²⁷ We continued publishing our CSER report throughout 2023.

- **Meta will supplement these reports with additional Australia-specific statistics, provided as part of this Annual Report process.**

While country-specific statistics should be interpreted with caution and have limitations in understanding misinformation and disinformation, Meta has again provided Australia-specific statistics in the spirit of transparency of our efforts.

²⁵ Meta, '[Meta Transparency Reports](#)', Meta Transparency Centre.

²⁶ Meta, '[Community Standards Enforcement Report](#)', Meta Transparency Centre.

²⁷ Meta, '[Community Standards Enforcement Report](#)', Meta Transparency Centre. 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>.

Several Australia specific metrics are included throughout this report including:

- The number of pieces of content removed across Facebook & Instagram in Australia for violating our Misinformation policies.
- The number of warnings displayed on distinct pieces of content on Facebook and Instagram in Australia (including reshares) based on articles written by our third-party fact-checking partners.
- The number of ads removed on Facebook and Instagram combined for violating our Misinformation policy.
- Specific metrics relevant to our media literacy campaigns conducted for the Aboriginal and Torres Strait Islander Voice referendum.

In addition to this, while we were unable to report some metrics for this reporting cycle, we are working to implement changes which will allow us to report on the reshare friction metric in Australia in future reports.

In relation to enforcement regarding Misinformation at an account level. As noted above, Meta's policies to tackle false claims about COVID-19 which could directly contribute to the risk of imminent physical harm changed in June 2023 following Meta's independent Oversight Board's advice.²⁸ We now only remove this content in countries with an active COVID-19 public health emergency declaration. This change has impacted our enforcement metrics on removals for this reporting period, but does not change our overall approach to fact-checking. These changes are an expected part of fluctuating content trends online.

²⁸ Oversight Board, '[Oversight Board publishes policy advisory opinion on the removal of COVID 19 Misinformation](#)', Oversight Board News, 20 April 2023.

Outcome 1e.

Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.

Signatories that provide services (other than search engines) whose primary purpose is to disseminate information to the public and which use recommender systems, commit to :

- A. make information available to end-users about how they work to prioritise information that end-users may access on these services; and
- B. provide end-users with options that relate to content suggested by recommender systems that are appropriate to the service.

- Meta will continue to provide greater transparency of our content ranking algorithms and give users more control over the content they see.

In June 2023, we launched Facebook System Cards²⁹ and Instagram System Cards³⁰ to help people to understand how AI shapes their product experiences and provides insights into how ranking systems dynamically work to deliver a personalised experience on Facebook and Instagram. These cards are available on Meta's Transparency Centre, and clicking into a card will bring up more information regarding how that specific AI system works, for example on Facebook Feed. These system cards were written in a way that can be understood by experts and non-experts alike.³¹

In addition to the detailed system cards, in June 2023, Meta also expanded the tools available to users to help them control the content they see and personalise their experience.³²

We also updated the 'Why am I seeing this ad' tool to provide more transparency about how activity both on and off of Facebook or Instagram inform machine learning models that Meta uses to shape the content that users see.³³ We also extended the 'Why am seeing this' tool for content from Feed, to Instagram Reels and Explore.

²⁹ Meta, '[Our approach to explaining ranking](#)', Meta Transparency Centre, 31 December 2023 and Meta, '[Building Generative AI features responsibly](#)', 27 September 2023.

³⁰ Meta, '[Our approach to explaining ranking](#)', Meta Transparency Centre, 31 December 2023 and Meta, '[Building Generative AI features responsibly](#)', 27 September 2023.

³¹ Meta, '[Our approach to explaining ranking](#)', Meta Transparency Centre, 31 December 2023 and Meta, '[Building Generative AI features responsibly](#)', 27 September 2023.

³² Meta, '[How AI Influences What You See on Facebook and Instagram](#)', Meta Newsroom, 29 June 2023

³³ Meta, '[Increasing our Ads Transparency](#)', Meta Newsroom, 14 February 2023.

We have also created centralised places on Facebook and Instagram where users can customise controls that influence the content they see on each app. Specifically:

- Users can visit their Feed Preferences on Facebook through the three-dot menu on relevant posts, as well as through Settings. For Instagram, users can visit the Suggested Content Control Center through the three-dot menu on relevant posts, as well as through Settings.
 - We also made the “Show more, Show less” feature,³⁴ which is available on all posts for Facebook and Instagram Feed and Reels, as well as Facebook Video, even more prominent via the three-dot-menu.
 - If a user does not want an algorithmically-ranked Feed, they can use the Feeds tab on Facebook. There is also an option to add people to the Favourites list on Facebook. Likewise, users on Instagram can change to a chronological feed, Following, based on the accounts they follow.
 - The Reduce tool allows users to adjust the degree to which we demote problematic or low-quality content in their Feed (our Content Distribution Guidelines outline some of the most significant reasons why problematic or low-quality content may receive reduced distribution in Feed).
- **Meta takes steps to limit the possible distribution of misinformation via recommendations.**

Meta takes steps to limit the possible distribution of misinformation via recommendations. Facebook and Instagram content that has been debunked³⁵ by non-partisan, third-party fact-checking organisations who partner with Meta, and have been certified by the International Fact-Checking Network (IFCN) is removed from recommendations.³⁶

Pages, Groups, Profiles, websites, and Instagram accounts that repeatedly share content rated False or Altered will be put under some restrictions for a given time period. This includes removing them from the recommendations we show people, reducing their distribution, removing their ability to monetise and advertise and removing their ability to register as a news Page.³⁷

³⁴ Meta, [‘The new AI-powered feature designed to improve Feed for everyone’](#), Meta Blog, 5 October 2022.

³⁵ Meta, [‘About fact-checking on Facebook, Instagram and Threads’](#), Meta Business Help Centre.

³⁶ Meta, [‘fact-checked Misinformation’](#), Meta Transparency Centre.

³⁷ Meta, [‘Penalties for sharing fact-checked content’](#), Meta Transparency Centre.

Outcome 2

Advertising and/or monetisation incentives for Disinformation are reduced.

Signatories will implement policies and processes that aim to disrupt advertising and/or monetisation incentives for Disinformation.

- **Meta sets a higher threshold for users to be able to advertise on our services, and takes action against users who spread misinformation.**

We have maintained the policy approach outlined in our 2021, 2022 and 2023 transparency reports.

We removed over 6,200 ads for violating our Misinformation policy from January to December 2023 in Australia. We also removed over 75,000 ads for not complying with SIEP ads policies³⁸ from January to December 2023 in Australia. This report is the first instance in which we've shared this data.

³⁸ Meta, ['Ads about Social Issues, Elections or Politics'](#), Meta Transparency Centre, 19 March 2024.

Outcome 3

The risk that Inauthentic User Behaviours undermine the integrity and security of services and products is reduced.

Signatories commit to take measures that prohibit or manage the types of user behaviours that are designed to undermine the security and integrity of their services and products, for example, the use of fake accounts or automated bots that are designed to propagate Disinformation.

Please see Outcome 1a for the actions Meta takes against inauthentic user behaviours.

Outcome 4

Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation.

Signatories will implement measures to enable users to make informed choices about news and factual information and to access alternative sources of information.

- **Meta provides contextual information around posts that users see from public Pages.**

With respect to these commitments, we have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports in terms of the contextual information around posts that users see from public Pages.

- **Meta provides a Climate Science Information Centre in Australia to connect users to authoritative information from leading climate organisations.**

Our platforms extend access to information and empower users to take action through tools like the Climate Info Finder Tool and the Climate Science Information Center, which launched in Australia in November 2021.³⁹ There are more than 18 million followers of the Center globally.⁴⁰

- **Meta uses in-product prompts to direct Australians to authoritative information on key topics.**

Consistent with previous reports under this code, we have launched in-app features that direct users to authoritative election information in Australia.

Ahead of the 2023 Aboriginal and Torres Strait Islander Voice Referendum, we developed reminder prompts to raise awareness about the referendum and link users to authoritative information.

More information is provided about these prompts in the Case Study: Aboriginal and Torres Strait Islander Voice Referendum in this report below.

³⁹ Meta, '[Our commitment to combatting climate change](#)', *Meta Newsroom*, 1 November 2021.

⁴⁰ Meta, '[Our approach to climate content](#)', 4 November 2022.

- **Meta promotes authoritative information by providing significant support to organisations such as the Australian, state and territory governments and First Nations health organisations to promote authoritative health information.**
- **Meta directs users to authoritative information when they search for high-priority topics on Facebook and Instagram.**
- **Meta directs users to authoritative information once they have seen or shared COVID-19 related misinformation.**

In response to **the three commitments above**, which focus on COVID-19-related information; as noted above, in June 2023 (in response to recommendations by Meta’s Oversight Board), we announced that our global COVID-19 misinformation rules would be altered as the global public health emergency declaration that triggered those rules had been lifted. We now only remove this content in countries with an active COVID-19 public health emergency declaration. In light of this policy change, we have shifted our time and resources to respond to new and emerging priorities, such as content governance relating to AI technology.

- **Meta will look for opportunities to continue to work with the Government on other ways to promote authoritative information.**

We have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports.

As an example of this, in July 2023, Meta facilitated a new training series called ‘Connect, Alert, Inform’, for emergency response organisations across Australia, New Zealand and the Pacific Islands. The training focused on helping to strengthen emergency communicators’ skills in using Meta’s services to build community and deliver critical disaster-related information.

The virtual training curriculum was developed in consultation with a disaster communications academic from the University of Technology Sydney, and in partnership with Emergency Management and Public Affairs (EMPA) and RMIT CrossCheck. It comprised sessions specifically tailored for emergency responders across a range of topics, including Meta’s crisis management tools; planning social media content; page moderation and account security; combating disaster related mis and dis-information; and social media advertising.

The training also included a separate simulation session, involving a fictional disaster scenario, where participants were encouraged to apply the learnings from the plenary training and consider how to best utilise social media as part of their emergency preparedness and response toolkit, and respond to challenges such as disaster-related misinformation and scams.

A diverse range of emergency communicators participated in the training, representing a mix of primarily government emergency response agencies, as well as law enforcement and humanitarian and disaster non-profit organisations from across the region.

- **Meta promotes public service announcements to our users to encourage them to be wary of potential misinformation.**

In terms of connecting Australians with authoritative information about key topics, in 2023, this was focused on the Aboriginal and Torres Strait Islander Voice Referendum, climate science information and COVID-19.

Ahead of the 2023 Aboriginal and Torres Strait Islander Voice Referendum, we developed reminder prompts to raise awareness about the referendum and link users to authoritative information.

You can find more information about these prompts in the Case Study: Aboriginal and Torres Strait Islander Voice Referendum.

Case study: Aboriginal and Torres Strait Islander Voice Referendum

In the lead up to Australia's Aboriginal and Torres Strait Islander Voice Referendum in October 2023, Meta developed a comprehensive strategy that focussed on proactively detecting and removing content that breaches our policies, combating misinformation, harmful content and CIB, and promoting civic participation.

Meta has been involved in more than 200 elections around the world since 2016, and we've learned key lessons from each one on combating misinformation, election interference and promoting civic participation. Alongside our routine integrity measures which we implemented for previous elections, we also rolled out additional measures to promote safety and integrity across our platforms in advance of the referendum.

This case study provides a comprehensive outline of our referendum integrity measures and their impact in Australia.

In the 2023 referendum campaign period, we developed referendum day reminders on Facebook and Instagram to encourage people to vote. Examples of these prompts for Facebook and Instagram, and stickers for Instagram can be found at the end of this case study.

These prompts reached a large number of Australians:

- Our Civic Products reached around 12.7 million users on Facebook with around 13.6 million impressions.
- For Instagram, around 6.55 million users were reached with around 6.57 million impressions.

We initiated a number of new programs of work in the lead up to the referendum in relation to misinformation and disinformation.⁴¹ This work includes:

- **Combating Misinformation and Foreign Interference Expanding Capacity for Meta’s Australian Fact-Checkers:** We know the importance of ensuring Australians had access to reliable information about the Aboriginal and Torres Strait Islander Voice referendum in Australia. Therefore, we provided a one-off funding boost to AAP and AFP, to increase their capacity in the lead up to the referendum. Our fact-checkers are independent and work to reduce the spread of misinformation across Meta’s services. When they rate something as false, we significantly reduce its distribution so fewer people see it. We also notify people who try to share something rated as false and add a warning label with a link to a debunking article.
- **Empowering People to Identify False News:** Since we know it’s not enough to just limit or remove harmful or misleading misinformation that people see, we launched a new media literacy campaign with Australian Associated Press,⁴² building on our “Check The Facts”⁴³ campaign which ran ahead of the 2022 Federal Election in October 2021 and early 2022. This shared tips and advice with people so that they could make informed decisions about what they read, trust and share. Further information about this campaign can be found below in the media literacy commitments under Outcome 6.
- **Combatting Influence Operations:** We have specialised global teams which worked during the referendum to identify and take action against threats to the elections and referendums, including signs of coordinated inauthentic behaviour across our apps. We also coordinated with the Government’s election integrity assurance taskforce and security agencies in the lead up to the referendum. We improved our AI so that we could more effectively detect and block fake accounts, which are often behind this activity.
- **Supporting academic analysis:** We supported analysis relating to information integrity surrounding the referendum debate and its outcomes, including RMIT CrossCheck’s work to promote accurate and corrective information on the

⁴¹ Meta, ‘[How Meta is preparing for the Voice to Parliament Referendum](#)’, *Medium*, 9 July 2023.

⁴² AAP, ‘[AAP FactCheck and Meta media literacy collab](#)’, AAP Press release, 28 August 2023.

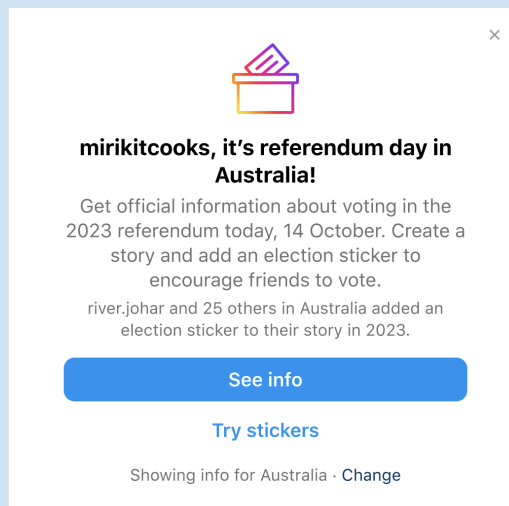
⁴³ AAP, ‘[Check the Facts](#)’, AAP Resources.

referendum debate and boost media literacy. We also supported La Trobe University's report 'Influencers and Messages: Analysing the 2023 Voice to Parliament Referendum Campaign',⁴⁴ which examined the main topics of the debate, key actors and campaign strategies for Yes and No, and the prevalence and influence of misinformation and disinformation.

Referendum day reminders used in the build up to and during the referendum on Facebook:



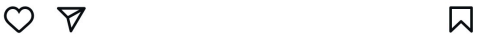
Referendum day reminders and stickers used in the build up to and during the referendum on Instagram:



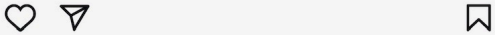
⁴⁴ Andrea Carson, Rebecca Strating, Simon Jackman, Max Grömping, Phoebe Hayman, Timothy B Gravelle, 'INFLUENCERS and MESSAGES: Analysing the 2023 Voice to Parliament Referendum Campaign', 17 April 2024.



Screenshots of selected videos from the Australian Associated Press media literacy campaign:



77 likes
aapfactcheck If you've ever encountered confusing or contradictory arguments, it may have been a deliberate attempt to mislead you. Incoherence is a common tactic used in disinformation, but being aware of it means you are less likely to fall for it.



40 likes
aapfactcheck Finding sources of information you can trust can be tricky, but good sources will always be open about the information they rely on, their processes, and their mistakes.

Outcome 5

Users are better informed about the source of Political Advertising

Signatories will develop and implement policies that provide users with greater transparency about the source of Political Advertising carried on digital platforms.

- **Meta requires all advertisers of political ads⁴⁵ to complete an ad authorisation, which includes verifying the advertiser’s identity.** We have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports.
- **Meta requires political ads to include a disclaimer disclosing who is paying for the ad.** We have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports.
- **Meta provides the Ad Library, a searchable archive of all political ads on our services in Australia.** We have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports.

In November 2023, Meta also announced a new AI Disclosure policy to help people understand when a social issue, election or political advertisement on Facebook or Instagram has been digitally created or altered, including through the use of AI.⁴⁶ Advertisers will disclose whenever a social issue, electoral, or political ad contains a photorealistic image or video, or realistic sounding audio, that was digitally created or altered to:

- Depict a real person as saying or doing something they did not say or do; or
- Depict a realistic-looking person that does not exist or a realistic-looking event that did not happen, or alter footage of a real event that happened; or
- Depict a realistic event that allegedly occurred, but that is not a true image, video, or audio recording of the event.

Meta will add information on the ad when an advertiser discloses in the advertising flow that the content is digitally created or altered. This information will also appear in the Ad Library.⁴⁷ If it is determined that an advertiser did not disclose as

⁴⁵ We define political ads as advertisements: (1) made by, on behalf of, or about a candidate for public office, a political figure, a political party or advocates for the outcome of an election to public office; or (2) about any election, referendum or ballot initiative, including "go out and vote" or election campaigns. We recognise the definition of Political Advertising in the voluntary industry code for disinformation and misinformation is broader than Facebook’s definition of “political ads”, as it also encompasses ads that we refer to as “social issue ads”.

⁴⁶ Meta, ‘Helping People Understand When AI Or Digital Methods Are Used In Political or Social Issue Ads’, Meta Newsroom, 8 November 2023.

⁴⁷ Meta, ‘[Meta Ad library](#)’.

required, Meta will reject the ad. Repeated failure to disclose may result in penalties against the advertiser.

- **Meta enables an Ad Library report that provides aggregated spend information about Pages undertaking political ads and social issue ads.**

We have maintained the approach outlined in our 2021, 2022 and 2023 transparency reports.

Outcome 6

Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation.

Signatories commit to support and encourage good faith independent efforts to research Disinformation and Misinformation both online and offline.

- **Meta supports research and events in relation to misinformation and media literacy.**

Since the 2023 transparency report, we have supported a number of events and research on misinformation, including:

- In the lead up to the 2023 Australian Aboriginal and Torres Strait Islander Voice Referendum, as mentioned above, we launched a new media literacy campaign in collaboration with the Australia Associated Press building on our “Check The Facts” campaign which ran ahead of the 2022 Federal Election in October 2021 and early 2022. This shared tips and advice with people so that they could make informed decisions about what they read, trust and share.⁴⁸
 - The objective of the campaign was to create engaging, organic-to-the-platform (Facebook and Instagram) content that would give Australian voting age adults the tools to understand what to look out for and what to fact check with the broad aim of limiting the influence misinformation in the period leading up to the referendum. The campaign videos were scripted and shot by AAP FactCheck journalists and aimed to generate awareness of tactics used to spread mis- and disinformation and by driving audiences to check the facts via the AAP Fact Check website.
 - The campaign ran for 6 weeks with a combined reach on Facebook and Instagram in Australia of over 10 million users, creating over 40 million impressions.⁴⁹
- Australian Media Literacy Alliance’s Media Literacy Summit in March 2023.⁵⁰
 - Meta sponsored and hosted the Australian Media Literacy Alliance’s Australian Media Literacy Summit, which brought together a range of journalists, academics, educators, librarians and other experts to discuss and learn about diverse ways to strengthen media literacy education in Australia.

⁴⁸ AAP, ‘[AAP FactCheck and Meta media literacy collab](#)’, AAP Press release, 28 August 2023.

⁴⁹ AAP reporting.

⁵⁰ Australian Media Literacy Alliance, ‘[Australian Media Literacy Summit](#)’, AMLA Event Press Release.

- Meta also supported the Western Sydney University’s 2023 Young People and News longitudinal survey
 - This survey provides findings about the news attitudes, practices and experiences of young Australians aged between 8-16 years.
- **Meta will continue to support research and events in relation to disinformation.**

For example, Meta has collaborated with the Australian Strategic Policy Institute (ASPI) for years, sharing our findings and insights about influence operations on our platforms and ingesting leads from networks they’ve identified.

- **Meta collaborates with researchers to undertake surveys of our users to assess their views on topics such as vaccines and climate change.**

Between March - May 2023 Meta conducted a ‘Pandemic Recovery’ Survey to inform pandemic response across the area of health education and economic recovery.⁵¹ The survey used aggregated indicators from self reported survey data in more than 21 countries for respondents aged 18 or over.

In 2023 Meta also conducted a ‘Climate Change Opinion’ Survey, in partnership with Yale, to explore public climate change knowledge, attitudes, policy preferences and behaviours.⁵² This survey included respondents from almost 200 countries and territories.

While Meta will continue to support research that has a positive social impact, our approach to COVID-19 has changed, following feedback from different stakeholders, including Meta’s independent Oversight Board. These changes reflect the changing nature and impact of COVID-19 globally since the start of the pandemic. As a result of this, future research topics may vary.

- **Meta provides data to researchers in a privacy-protective way via the Facebook Open Research and Transparency (FORT) initiative.**

The FORT initiative experienced a name change and all relevant transparency efforts can be found on our Transparency Centre.⁵³ Specifically, we have research tools called Meta Content Library and API that share publicly accessible data from Facebook and Instagram. The MCL API is accessible in a secure clean room environment. Meta will make the Meta Content Library and API research tools available to third-party fact-checking partners and qualified users.

⁵¹ Meta, [‘Pandemic Recovery Survey’](#), Meta Data for Good Centre.

⁵² Meta, [‘Climate Change Opinion Survey’](#), Meta Data for Good Centre.

⁵³ Meta, [‘Research Tools and Data Sets’](#), Meta Transparency Centre.

Outcome 7

The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.

All Signatories will make and publish the annual report information in section 7

- Meta will continue to publish annual reports in Australia, such as these, to be transparent about the steps we are taking to combat disinformation and misinformation.

We met this requirement in 2021, 2022 and 2023 by publishing a transparency report on our Meta Australia blog, and speaking publicly to the media about our work.⁵⁴

The publication of this report satisfies this requirement for 2024.

⁵⁴ Meta, '[Facebook's response to Australia's disinformation and misinformation code](#)', *Meta Australia Blog*, 21 May 2021.

Appendix A: 2023 specific commitments made by Meta under the industry code on disinformation and misinformation

The 38 commitments for 2023 are outlined below:

Outcome 1: Combatting misinformation / disinformation	Outcome 1a. Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures. <ul style="list-style-type: none">• Meta removes networks of accounts, Pages and Groups that violate our inauthentic behaviour policy, including disinformation that violates our policy on Inauthentic Behaviour.• Meta provides transparency about accounts, Pages and Groups removed under our Inauthentic Behaviour policy.• Meta partners with experts and organisations who assist in providing tips or further investigation about possible inauthentic behaviour on our services.• Meta removes misinformation that violates the misinformation section of our Community Standards, and will review its Misinformation and Harm Policy, in line with recommendations from the Oversight Board.• Meta removes manipulated videos, also known as “deepfakes”, that violates our Manipulated Media policy.• Meta removes election-related misinformation that may constitute voter fraud or interference under our Coordinating Harm and Promoting Crime policy.• Meta removes fake accounts.• Meta allows for appeals in instances where users may disagree with our enforcement, including to the independent and external Oversight Board.• Meta partners with third-party fact-checking organisations, globally and in Australia, to assess the accuracy of content on our services.• Meta applies a warning label to content found to be false by third-party fact-checking organisations.
--	---

- Meta reduces the distribution of content found to be false by third-party fact-checking organisations.
- Meta proactively searches for content that matches content debunked by our fact-checking partners, to apply the same treatments.
- Meta takes action on Pages, Groups, accounts or websites found to repeatedly share misinformation.

Outcome 1b. Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.

- Meta makes available a detailed list of claims that we consider to violate our COVID-19 Misinformation & Harm policy.
- Meta makes information available via a dedicated website that outlines our efforts to combat misinformation.

Outcome 1c . Users can report content and behaviours to Signatories that violate their policies under 5.10 through publicly available and accessible reporting tools.

- Meta makes on-platform reporting channels available to users for false information.

Outcome 1d. Users will be able to access general information about Signatories' actions in response to reports made under 5.11.

- Meta makes global transparency reports available regularly.
- Meta will supplement these reports with additional Australia-specific statistics, provided as part of this Annual Report process.

[*NEW*] Outcome 1e. Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.

- Meta will continue to provide greater transparency of our content ranking algorithms and give users more control over the content they see.

	<ul style="list-style-type: none"> • Meta takes steps to limit the possible distribution of misinformation via recommendations.
<p>Outcome 2: Disrupt monetisation and advertising incentives</p>	<ul style="list-style-type: none"> • Meta sets a higher threshold for users to be able to advertise on our services, and takes action against users who spread misinformation.
<p>Outcome 3: Combat inauthentic user behaviour</p>	<p>See items listed under Outcome 1.</p>
<p>Outcome 4: Empower consumers to be informed</p>	<ul style="list-style-type: none"> • Meta provides contextual information around posts that users see from public Pages. • Meta provides a Climate Science Information Centre in Australia to connect users to authoritative information from leading climate organisations. • Meta uses in-product prompts to direct Australians to authoritative information on key topics. • Meta promotes authoritative information by providing significant support to organisations such as the Australian, state and territory governments and to promote authoritative health information. • Meta directs users to authoritative information when they search for high-priority topics on Facebook and Instagram. • Meta directs users to authoritative information once they have seen or shared COVID-19 related misinformation. • Meta will look for opportunities to continue to work with the Government on other ways to promote authoritative information. • Meta promotes public service announcements to our users to encourage them to be wary of potential misinformation.

**Outcome 5:
Political
advertising**

- Meta requires all advertisers of political and social issue ads to complete an ad authorisation, which includes verifying the advertiser’s identity.
- Meta requires political and social issue ads to include a disclaimer disclosing who is paying for the ad.
- Meta provides the Ad Library, a searchable archive of all political and social issue ads on our services in Australia.
- Meta enables an Ad Library report that provides aggregated spend information about Pages undertaking political and social issue ads.

**Outcome 6:
Research**

- Meta supports research and events in relation to misinformation and media literacy.
- Meta supports research and events in relation to disinformation.
- Meta collaborates with researchers to undertake surveys of our users to assess their views on topics such as vaccines and climate change.
- Meta provides data to researchers in a privacy-protective way via the Facebook Open Research and Transparency initiative.

**Outcome 7:
Annual reports**

- Meta will continue to publish annual reports in Australia, such as these, to be transparent about the steps we are taking to combat disinformation and misinformation.



Australian Code of Practice on Disinformation and Misinformation

Microsoft and LinkedIn Annual Transparency Report May 2024

Summary

Microsoft is pleased to file this report on our commitments under the voluntary Australian Code of Practice on Disinformation and Misinformation (the **Code**), covering the reporting period of calendar year 2023.

We have submitted Transparency Reports under the Code every year since 2021. In each Transparency Report, we have shown how Microsoft is committed to instilling trust and security across our products and services, and across the broader online ecosystem. We continue to recognise that fighting disinformation is a key element to creating a trustworthy and safe online environment and continue to increase our efforts to counter these threats.

We also recognise that there is not a one size fits all approach to this work, and instead there needs to be a whole of society strategy that recognises that not all people or platforms are the same and that different measures may be more effective than others in improving the information environment.

The Microsoft services in scope of the Code are:

- **Microsoft Advertising:** Microsoft's proprietary online advertising network, which serves most ads displayed on Bing Search and provides advertising to most other Microsoft services that display ads.
- **Bing Search:** a web search engine which provides a variety of services including web, video, image, and map search products. Bing Search does not host the content appearing in search results, does not control the operation or design of the indexed websites and has no ability to control what indexed websites publish. This reporting period also discusses Copilot in Bing, which was released in February 2023 as 'Bing Chat' and is now marketed separately under the Copilot brand.
- **Microsoft Start:** a service which delivers licenced news and content across web and mobile on behalf of Microsoft customers and syndication partners.
- **LinkedIn:** a real identity online networking service for professionals to connect and interact with other professionals, to grow their professional network and brand, and to seek career development opportunities. It operates via websites and mobile apps and includes user-generated content.

Our approach

Microsoft announced the first [Information Integrity Principles](#) in 2022. These principles continue to be adopted across all impacted Microsoft products and teams to ensure an enterprise approach to information integrity while also recognising the immense diversity across the company. The four information integrity principles are:

- **Freedom of Expression:** We will respect freedom of expression and uphold our customers' ability to create, publish, and search for information via our platforms, products, and services.



- **Authoritative Content:** We will prioritise surfacing content to counter foreign cyber influence operations by utilising internal and trusted third-party data on our products.
- **Demonetisation:** We will not wilfully profit from foreign cyber influence content or actors.
- **Proactive Efforts:** We will proactively work to prevent our platforms and products from being used to amplify foreign cyber influence sites and content.

Since our last Transparency Report, the focus on artificial intelligence (**AI**) and interest in understanding how AI could affect the spread of disinformation has continued to grow. While AI certainly poses challenges in the information integrity space, we also see many opportunities for AI to assist and streamline defenders' work in detecting and assessing influence operations. To be clear, challenges include the evolving tactics and potential efforts to create or disseminate malicious content. However, Microsoft is fully committed to utilizing best in class tools, practices, and technology to help mitigate the risks of its services being used to further disinformation.

Serving as a leader in AI research, we are committed to proactively publicize our threat detection efforts for the benefit of society. As such, we have adopted six focus areas to combat the harmful use of deceptive AI:

1. A strong safety architecture
2. Durable media provenance and watermarking
3. Safeguarding our services from abusive content and conduct
4. Robust collaboration across industry and with governments and civil society
5. Modernized legislation to protect people from the abuse of technology
6. Public awareness and education

Relatedly, in February 2024, Microsoft and LinkedIn were two of 20 companies that announced a new [Tech Accord to Combat Deceptive Use of AI in 2024 Elections](#). The goal is straightforward but critical – to combat video, audio, and images that fake or alter the appearance, voice, or actions of political candidates, election officials, and other key stakeholders or that provide false information to voters about when, where, and how they can lawfully vote.

Additionally, Microsoft is an active member of the [Coalition for Content Provenance and Authority \(C2PA\)](#), working to further define and standardize an end-to-end process for



publishing, distribution, and attaching signals to a piece of content to demonstrate its integrity.

These examples form just one part of Microsoft's **Democracy Forward** initiative, a coordinated program launched back in 2018 to coordinate and track the work undertaken across the company on protecting and strengthening democratic institutions.

In Australia, Microsoft is an inaugural signatory to the Electoral Council of Australia and New Zealand (ECANZ) Statement of Intent with Online Platforms, designed to support Australian electoral management bodies and online platforms to work together to promote and support the integrity of electoral events in Australia. During the reporting period, Microsoft worked closely with both the Australian Electoral Commission and the Victorian Electoral Commission to establish arrangements to manage content referrals related to breaches of the relevant electoral laws in relation to the Aboriginal and Torres Strait Islander Voice to Parliament Referendum (the **Voice Referendum**). Neither Microsoft Advertising nor LinkedIn accept political advertising.

Microsoft Advertising

Microsoft Advertising works both with advertisers, who provide it with advertising content, and publishers, such as Bing Search, who display these advertisements on their services. Microsoft Advertising employs a distinct set of policies and enforcement measures with respect to each of these two categories of business partners to prevent the spread of disinformation through advertising.

Bing Search

Bing Search is an online search engine with the primary objective of connecting users with the most relevant search results from the web. Users come to Bing with a specific research topic in mind and expect Bing to provide links to the most relevant and authoritative third-party websites on the internet that are responsive to their search terms. Therefore, addressing misinformation in organic search results often requires a different approach than may be appropriate for other types of online services. Blocking content in organic search results based solely on the truth or falsity of the content can raise significant concerns relating to fundamental rights of freedom of expression and the freedom to receive and impart information.

While Bing's efforts may on occasion involve removal of content from search results (where legal or policy considerations warrant removal), in many cases, Bing has found that actions such as targeted ranking interventions, or additional digital literacy features such as Answers pointing to high authority sources, trustworthiness signals, or content provenance indicators, are more effective. Bing regularly reviews the efficacy of its measures to identify additional areas for improvement and works with internal and external subject matter experts in key policy areas to identify new threat vectors or improved mechanisms to help prevent users from being unexpectedly exposed to harmful content in search results that they did not expressly seek to find.



Copilot in Bing provides a next-generation search experience for users to find the web content they are seeking more efficiently, including through more conversational questions and interactions with the service. It is built on longstanding safety systems underpinning Bing search, supplemented by additional protections for new risks related to AI. Microsoft has partnered closely with Microsoft's Responsible AI team to proactively address these harms and has been transparent about its approach in [Copilot in Bing: Our approach to Responsible AI](#). Bing continues to evolve these features based on user and external stakeholder feedback.

LinkedIn

LinkedIn is a real identity online networking service for professionals to connect and interact with other professionals, to grow their professional network and brand, and to seek career development opportunities.

LinkedIn is part of its members' professional identity and has a specific purpose. Activity on the platform and content members share can be seen by current and future employers, colleagues, potential business partners and recruitment firms, among others. Given this audience, members largely limit their activity to professional areas of interest and expect the content they see to be professional in nature.

LinkedIn is committed to keeping its platform safe, trusted, and professional, and respects the laws that apply to its services. On joining LinkedIn, members agree to abide by LinkedIn's [User Agreement](#) and its [Professional Community Policies](#), which expressly prohibit the posting of information that is intentionally deceptive or misleading.

When LinkedIn sees content or behaviour that violates its Professional Community Policies, it takes action, including the removal of content or the restriction of an account for repeated abusive behaviour. In 2023, LinkedIn globally blocked more than 120 million fake accounts (a majority of which were stopped at registration) and removed more than 139,009 pieces of misinformation. Over the same period, LinkedIn blocked more than 600,000 fake accounts attributed to Australia and removed 1,540 pieces of misinformation reported, posted, or shared by Australian members.

Microsoft Start

Microsoft Start is a personalised feed of news and informational content from publishers available in a number of Microsoft products, including a standalone website (MSN.com), a mobile app on both Android and iOS, the News and Interests experience on the Windows 10 taskbar, the Widgets experience in Windows 11, and the Microsoft Edge new tab page. On Microsoft Start, we have policies to specifically address disinformation and misinformation on clear and well-defined misinformation narratives.

Commitments under the Code

Commitment	Relevant Microsoft service
1a: Contribute to reducing risk of harm by adopting scalable measures	Bing Search, Microsoft Start, Microsoft Advertising, LinkedIn
1b: Users informed about types of behaviours and content prohibited/managed	Microsoft Start, Microsoft Advertising, LinkedIn
1c: Users can report content that violates policy through accessible reporting tools	Bing Search, Microsoft Start, Microsoft Advertising, LinkedIn
1d: Users can access general information about response	Bing Search, Microsoft Start, Microsoft Advertising, LinkedIn
1e: Users will be able to access general information about use of recommender systems and have options related to content suggested by recommender systems	LinkedIn
2: Advertising and/or monetisation incentives reduced	Microsoft Advertising, LinkedIn
3: Risk of inauthentic behaviours undermining integrity and security of services/products reduced	Bing Search, Microsoft Advertising, LinkedIn
4: Users are enabled to make informed choices about sources of news and factual content and to identify misinformation	Bing Search, Microsoft Start, LinkedIn
5: Users better informed about source of Political Advertising	Microsoft Advertising, LinkedIn
6: Support efforts of independent research	Microsoft
7: Public access to measures to combat disinformation and misinformation	Bing Search, Microsoft Start, Microsoft Advertising, LinkedIn

Unless otherwise specified, data provided in this Transparency Report is for 2023 calendar year.

Reporting Against Commitments

Objective 1: Safeguards against Disinformation and Misinformation

Outcome 1a: Signatories contribute to reducing the risk of harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures.

Microsoft reduces the risk of harms that may arise from the propagation of disinformation and misinformation on **Bing Search, Microsoft Start, Microsoft Advertising** and **LinkedIn** through the application of our internal policies and scalable measures.

(O.1a) Bing Search

Bing Search is an online search engine that provides a searchable index of websites available on the internet. Bing Search does not have a news feed for users, allow users to post and share content, or otherwise enable content to go “viral” on its service. Nonetheless, disinformation could at times appear in both organic search results, and we take active steps to counter it. As emphasised in the summary above, addressing disinformation in organic search results often requires a different approach than may be appropriate for other types of online services, such as social media services.

Bing Search’s primary mechanism for combatting misinformation in search is via ranking improvements that take into account the quality and credibility (**QC**) of a website and work to rank higher quality and more authoritative pages over lower authority content. Bing Search describes the main parameters of its ranking systems, including QC, in depth in [How Bing Delivers Search Results](#). Abusive techniques and examples of prohibited SEO activities are described in more detail in the [Bing Webmaster Guidelines](#).

Determining the QC of a website includes evaluating the clarity of purpose of the site, its usability, and presentation. QC also consists of an evaluation of the page’s “authority”, which includes factors such as:

- **Reputation:** What types of other websites link to the site? A well-known news site is considered to have a higher reputation than a brand-new blog.
- **Level of discourse:** Is the purpose of the content solely to cause harm to individuals or groups of people? For example, a site that promotes violence or resorts to name-calling or bullying will be considered to have a low level of discourse, and therefore lower authority, than a balanced news article.
- **Level of distortion:** How well does the site differentiate fact from opinion? A site that is clearly labelled as satire or parody will have more authority than one that tries to obscure its intent.

- **Origination and transparency of the ownership:** Is the site reporting first-hand information, or does it summarize or republish content from others? If the site doesn't publish original content, do they attribute the source?

Bing Search's general spam policies also prohibit certain practices intended to manipulate or deceive the Bing Search algorithm, including techniques employed by malicious actors in the spread of misinformation. Bing's spam policies are detailed in the "Abuse and Examples of Things to Avoid" section of the Bing Webmaster Guidelines.

Although the Bing Search algorithm endeavours to prioritise relevance, quality, and credibility in all scenarios, in some cases Bing Search identifies a threat that undermines the efficacy of its algorithms. When this happens, Bing Search employs "defensive search" strategies and interventions to counteract threats.

Defensive search interventions may include:

- algorithmic interventions (such as quality and credibility boosts or demotions of a website);
- restricting autosuggest or related search terms to avoid directing users to potentially problematic queries; and
- manual interventions for individual reported issues or broader areas more prone to misinformation or disinformation (e.g., elections, pharmaceutical drugs, or COVID-19).

Defensive Search Interventions, Australia

	January – December 2022		January – December 2023	
	Queries	Impressions	Queries	Impressions
Total	64,104	4,441,099	136,450	2,270,775
Ukraine related[#]	45,100	1,072,939	17,964	618,366

[#]Ukraine data from February to December 2022 for the 2022 reporting year.

Bing regularly partners with independent third-party organisations to obtain threat intelligence on emerging narratives and mis/disinformation patterns and tactics that helps to inform potential algorithmic interventions. Bing Search also takes action to remove auto-suggest and related search terms that could inadvertently result in problematic or misleading content. Bing Search also may include answers or public service announcements at the top of search results pointing users to high authority information on a searched topic or



warnings on particular URLs known to contain harmful information (such as unaccredited online pharmacies and sites containing malware).

While Bing Search generally strives to rank its organic search results so that trusted, authoritative news and information appear first and provides tools that help Bing Search users evaluate the trustworthiness of certain sites, we also believe that enabling users to find all types of information through a search engine can provide important public benefits. Bing users have many legitimate reasons for seeking out content in search that may be harmful or offensive in other contexts (such as for research purposes) and unduly restricting access to information can pose risks to users to fundamental rights.

Generative AI Features (Copilot in Bing)

In February 2023, Microsoft launched new generative AI features, including an AI-enhanced web search experience that allows users to quickly and easily obtain answers through a chat experience that is 'grounded' in high authority web search results (meaning chat answers include links to supporting webpages). These AI features were originally known as Bing Chat and Image Creator, but since have undergone continued improvements and have evolved into a new distinct family of AI services under the Microsoft Copilot brand. One such endpoint is in Bing Search, which integrates Copilot functionality to provide users with a modern, natural language-based search interface ("Copilot in Bing").

Copilot in Bing combines traditional search functionality with the capabilities of large language models. This enables users to ask more complex and nuanced questions and engage in more natural conversations on search topics. Copilot in Bing uses AI to concisely summarize relevant information in search results, with links to supporting webpages where users can continue their research and evaluate the credibility of cited resources. Users can also use Copilot to generate creative content, such as code, poems, jokes, stories, and images.

Copilot in Bing's primary functionality is, like traditional Bing search, to provide users with links to third party content responsive to their search queries. As such, the ranking algorithms and spam/abuse policies described above continue to be Bing's key defence against manipulation and abuse. However, in recognition of the potential risks that these new technologies can pose, Microsoft also implements interventions designed specifically to address manipulation and risks of misinformation in generative AI features. As a result, Microsoft has supplemented its existing threat identification and mitigation processes with additional risk assessment and mitigation processes based on [Microsoft's Responsible AI program](#). Guided by our [Responsible AI Standard](#), we sought to identify, measure, and mitigate potential harms and misuse of new Bing generative AI experiences while securing the transformative and beneficial uses that the new experience provides.

How Microsoft limits the impact of “hallucinations” on Copilot in Bing

Microsoft has adopted several measures to help ensure that Copilot in Bing relies on reliable sources of information to limit the impact of “hallucinations” that may appear in outputs.

- **Grounding of outputs.** Copilot in Bing’s responses to user prompts, when the user asks for factual information, are grounded in web search results. This means they are based on the same ranking algorithms and safety infrastructure that Microsoft applies to traditional web search in Bing (described above, and in [How Bing Delivers Search Results](#)). As part of the grounding process, Copilot in Bing outputs to such queries include footnotes to the third-party sources from which they are drawn and provide links to these sources so that users can navigate directly to them and independently evaluate their credibility and reliability, just as they do with traditional web searches, and assess the accuracy of the output summary of that source material.
- **Metaprompts, filters, and classifiers.** Metaprompts are essentially instructions that Microsoft adds to guide system behaviour and tailor its output so that the system behaves in accordance with Microsoft’s AI Principles and user expectations, including to prevent the generation of responses that could be harmful to the user, to avoid giving users access to underlying safety instructions that could allow them to bypass safety protections, and to provide disclaimers in answers where there is uncertainty as to whether the user could be harmed by the results generated. As another layer of protection, Microsoft has implemented additional filtering and classifiers to prevent Copilot in Bing responses from returning harmful content to users, including in some cases to block users from generating responses based on prompts that are likely to violate our Code of Conduct or to prevent the service from returning low-authority web materials.
- **Reliance on third-party signals.** Microsoft supports credibility ratings that reputable third-party organisations apply to online news and other sources. We discuss these mechanisms in detail in Part O.4 of our report, below. Because outputs generated by Copilot in Bing are grounded in Bing’s web search results, these outputs likewise benefit from these signals, thereby providing independent, third-party sources through which users can assess the credibility of the sources underpinning Copilot in Bing outputs. Users that have enabled the NewsGuard browser plugin also see NewsGuard reliability ratings and nutrition labels in responses generated in Copilot in Bing, which can further help users evaluate the reliability of sources cited in these responses.
- **Informing users they are interacting with AI.** Microsoft’s RAI Standard requires Microsoft to design its AI systems to “inform people that they are interacting with an AI system or are using a system that generates or manipulates image, audio, or video content that could falsely appear to be authentic.” Consistent with these requirements, Microsoft has taken a multifaceted approach to ensure that users of Copilot in Bing are aware that the outputs it produces are not generated by a human and might not be accurate, as well as reminders to check the veracity of content provided by Copilot

in Bing. These steps include: (1) stating at the top of the Copilot in Bing web page that “Bing is your AI-powered copilot for the web” (or similar language), thereby disclosing to users that these responses are generated by AI; (2) also stating on the page that “Copilot uses AI. Check for mistakes” (or similar language), thereby signalling to users that they should not assume that responses are accurate; (3) defining conversational experiences in the [Terms of Use](#) by reference to AI-powered generative experiences; and (4) including citations and links in the responses themselves to the web sources from which the response was derived, thereby alerting users to the source of the information and enabling them to learn more by clicking on the links.

Microsoft has worked continuously to improve and adjust safety mitigations, policies, and user experiences within Copilot in Bing to minimize the risk they may be used for deceptive or otherwise problematic purposes in violation of Microsoft policies and regularly evaluates and improves safety measures. Additional detail on how Microsoft approached responsible AI in the development of Copilot in Bing is available in [Copilot in Bing: our Approach to Responsible AI](#).

We note that Bing does not host user content and users cannot post or share content directly on the Bing service, including Copilot in Bing. In addition, Microsoft undertakes specific mitigations to address the risks that individuals may attempt to use generative AI to create deep fakes or manipulated media to spread misinformation. Although Bing does not have the ability to monitor third party platforms for publication of content created through Bing’s services, Bing has implemented safeguards to help to minimize the risk that bad actors can use Bing generative AI experiences to create mis/disinformation that could potentially be shared on other platforms. See more [here](#), [here](#) and [here](#).

(O.1a) Microsoft Start

Microsoft Start delivers high-quality news across web and mobile experiences for Microsoft as well as a growing number of syndication partners. Microsoft Start’s model reduces risk of disinformation and misinformation being propagated. Misinformation in our licenced content feed has been exceedingly rare.

- Our content providers are vetted and must adhere to a strict set of standards that prohibit false information, propaganda and deliberate misinformation.
- Microsoft Start is free to download, with no limits on number of articles or videos a user can view.

Microsoft Start Community supports diverse, authentic conversations and content about issues and events. Our [Community Guidelines](#) are designed to uphold these values and we strive to provide transparency and clear guidance on how to comply with them.

- If a contribution is flagged, it will be reviewed. If it does not meet the community guidelines it will be removed.



- User activity feed shows if any comments have been removed and users are able to appeal the decision.

When necessary, Microsoft Start will suspend a user’s ability to comment. Continued refusal to meet standards may result in permanent ban, which users have an opportunity to appeal.

A misinformation trait will define what can and cannot be said on our platform about a particular topic. We have specific policies for managing misinformation relating to well-defined misinformation narratives with potential for real-world harm - which includes disabling comments for certain articles to reduce propagation of disinformation.

- In response to the ongoing invasion of Ukraine, we maintain a corresponding misinformation trait to prevent the Microsoft Start Community from becoming a platform for disinformation in relation to this conflict.
- In response to the Israel-Hamas conflict, we have disabled comments on a significant quantity of associated articles and videos as a proactive action against propagation of mis/disinformation through comments.
- Comments were also disabled on articles and videos relating to events taking place in Australia, such as the Voice Referendum and certain high-profile sexual assault and defamation cases.

Microsoft Start saw a dramatic decline in misinformation for this reporting period.

In this reporting period, we saw a dramatic decline in misinformation numbers, which can be attributed to several factors.

Firstly, the misinformation traits that we track (COVID-19, QAnon and the Russian-Ukraine conflict) experienced dips of varying degrees in traffic from previous reporting periods.

Secondly, Microsoft Start began integrating GPT4 enabled content moderation solutions for comments starting mid-way through the reporting period. The data we track and present in these reports relates to user-initiated complaint and takedown processes, whereas the GPT4 solution monitors and removes content automatically. Consequently, there are less violative comments which users could manually report.

In the reporting period, 1,496,208 comments were proactively blocked in Australia by these systems on Microsoft Start.

Microsoft Start Comments, Australia Takedowns, October 2021 – December 2022

	October – December 2021[#]		January - December 2022		January – December 2023	
Total takedowns	73,700	100%	849,000	100%	9955	100%

Misinformation – all*	1,899	2.5%	9,256	1.09%	49	0.49%
Misinformation – COVID-19	1,810	2.4%	8,655	1.01%	33	0.33%
Misinformation - QAnon	89	0.12%	425	0.05%	4	0.04%
Misinformation - Russia/Ukraine [^]			128	0.01%	12	0.12%

#Comments data prior to October 2021 is not a reliable metric as the function was only in its early stages.

* Misinformation total includes comments which have more than one trait labelled; percentages are rounded; sub-category list is not exhaustive.

[^]Russia/Ukraine misinformation trait was introduced in February 2022.

(O.1a) Microsoft Advertising

Microsoft Advertising's [Misleading Content Policies](#) prohibit advertising content that is misleading, deceptive, fraudulent, or that can be harmful to its users, including advertisements that contain unsubstantiated claims, or that falsely claim or imply endorsements or affiliations with third party products, services, governmental entities, or organisations.

In 2023, Microsoft Advertising took a number of actions to ensure a safe and trusted experience.

This included:

- taking down more than 8 billion ads and product offers for various policy violations. We suspended nearly 537,000 customers and blocked ~372,500 ads with websites that either contain content not allowed in our policy or spread disinformation;
- making use of significant advancements in artificial intelligence (AI) to quickly adapt to new patterns and methods used by bad actors;
- ensuring that our protection mechanism involved coverage for all types of content such as text, images, and videos to quickly detect malicious activity in our system;
- making advancements in our human moderation workflows to capture more insights from reviews, continuously improving our systems;
- leveraging intelligent tools to allow our human reviewers to establish linkages between various accounts and discover fraud rings quickly and efficiently;
- developing automated detection mechanisms to enforce new policies on information integrity, including developing new logic in the system to prevent receiving requests to show ads on web domains that may violate our disinformation policies; and,
- further iterating those automated detection mechanisms, including new automated classifiers to detect misleading claims relating to false information and consumer scams, such as financial scams, unsupported pricing claims and sensationalized ads, and misleading celebrity endorsements.

Microsoft Advertising deploys a range of policy-based and proactive measures to reduce the risk of harms associated with disinformation and misinformation, including:

- Our [Relevance and Quality Policies](#), which manage the relevancy and quality of the advertisements that it serves through its advertising network. These policies deter advertisers from luring users onto sites using questionable or misleading tactics (e.g., by prohibiting advertisements that lead users to sites that misrepresent the origin or intent of their content).
- Our [Sensitive Advertising Policies](#), Microsoft reserves the right to remove or limit advertising permanently or for a period of time in response to a sensitive tragedy, disaster, death or high-profile news event, particularly if the advertising may appear to exploit events for commercial gain or may affect user safety.



Just prior to the start of this reporting period, in December 2022, Microsoft Advertising rolled out revised network-wide policies to avoid the publishing and carriage of harmful disinformation and the placement of advertising next to disinformation content. Such policies prohibit ads or sites that contain or lead to disinformation. Our policy states, “We may use a combination of internal signals and trusted third-party data or information sources to reject, block, or take down ads or sites that contain disinformation or send traffic to pages containing disinformation. We may block at the domain level landing pages or sites that violate this policy.” See our [main policy page](#).

As previously reported, Microsoft Advertising is continuing to prevent serving advertising related to the Russia-Ukraine conflict, pursuant to its Sensitive Advertising Policies. Relatedly, Microsoft Advertising is preventing serving advertising related to the Israel-Hamas conflict pursuant to its Sensitive Advertising Policies. Under this policy, Microsoft Advertising reserves the right to remove or limit advertising in response to a sensitive or high-profile news event to prevent the commercial exploitation of such events and to ensure user safety.

Microsoft Advertising Global Ad Takedowns

2020	2021	2022	2023
1.6 billion	3 billion	7.2 billion	8.2 billion

Microsoft Advertising Ad Safety in Australia

*Although not possible to estimate with precision, the year-over-year growth in the number of rejections and related figures may be due to the expansion of Microsoft Advertising in new international markets, and the growth in adoption of certain advertising formats compared to the previous year.

Action	2021		2022		2023	
	Global	Australia	Global	Australia	Global	Australia
Rejections	3b	191m	7.2b	1b	8.2b	1.33b
Total appeals	72,413	7,025	127,158	14,536	132,910	7,747
Total appeals overturned	28,965	3,248	101,537	9,522	95,738	6,858
Total complaints	70,000	201	35,667	285	46,168	1,090

Complaint: Policy violation	20,934	68	1,156	57	1,411	14
Complaint: Trademark infringement	34,700	127	32,213	153	31,223	238
Complaint: User safety issues	416	5	1,805	58	13,533	838
Complaints: Other	13,950	69	493	17	407	7
Total entity takedowns	250,124	2,956	551,424	118,321	1,746,324	332,332
Average processing time	~36 hours	~36 hours	~36 hours	~36 hours	~36 hours	~36 hours

Microsoft's Advertiser Identify Verification Program

[The Advertiser Identity Verification](#) program, designed to verify the identity of the advertisers who buy ads through Microsoft Advertising, is available across our ad network, including in Australia. In 2023, 41,940 accounts opted for AIV verification in Australia and out of these, 41,592 accounts were successfully verified. The system enables customers to see ads from trusted sources. The selected advertisers are required to establish their identity as a business or as an individual by submitting all necessary information and documents.

Microsoft ensures that all advertisements on our services are clearly distinguishable from editorial or other non-sponsored content.

- All Microsoft services that display ads served by Microsoft Advertising clearly distinguish sponsored from non-sponsored content by displaying an advertising label in a readily noticeable location on the page. An example of how ads are displayed is shown in red below. Clicking on the information icon or downward arrow next to an advertising label displays a click through to the [ad setting page](#).

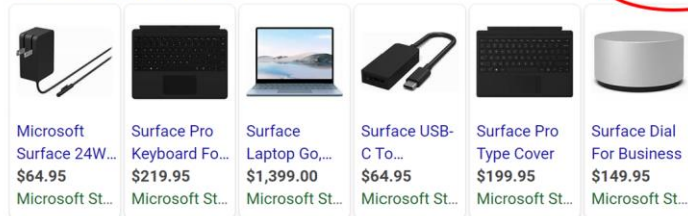
Surface Devices, Accessories - Microsoft Store

<https://www.microsoft.com/en-au/store/collections/surfacelist>

The most portable Surface touchscreen 2-in-1 is perfect for your everyday tasks, homework and play. Designed to light up the best of Windows 11, Surface Go 3 is optimised for digital pen and...

See microsoft surface

Ads ⓘ



Microsoft Advertising similarly requires all its publishers to use a clear and prominent label indicating that the advertisements served by Microsoft Advertising on their properties are sponsored. Microsoft Advertising proactively reviews publisher partners to enforce this requirement.

(O.1a) LinkedIn

To help keep **LinkedIn** safe, trusted, and professional, our [Professional Community Policies](#) clearly detail the range of objectionable and harmful content that is not allowed on LinkedIn. Fake accounts, misinformation, and inauthentic content are not allowed, and we take active steps to remove it from our platform.

LinkedIn has automated defences to identify and prevent abuse, including inauthentic behaviour, such as spam, phishing and scams, duplicate accounts, fake accounts, and misinformation. Our Trust and Safety teams work every day to identify and restrict inauthentic activity. We're regularly rolling out scalable [technologies](#) like machine learning models to keep our platform safe.

Using the process described in response to Outcome 1c below, LinkedIn members also can report content they believe violates our Professional Community Policies, including misinformation, inauthentic content, and fake accounts. If reported or flagged content violates the Professional Community Policies, it will be removed from the platform. We may also restrict the offending member's LinkedIn account, depending on the severity of the violation and any history of abuse.

LinkedIn has numerous workstreams that address misinformation, particularly during crisis situations. For instance, LinkedIn's in-house editorial team provides members with trustworthy content regarding global events, including Russia's war against Ukraine and the Israel-Hamas conflict. LinkedIn has an internal team of hundreds of content reviewers located all over the world providing 24/7 coverage and includes specialists in a number of languages.

The [LinkedIn Community Report](#) describes actions we take on content that violates our Professional Community Policies and User Agreement. It is published twice per year and



covers the global detection of fake accounts, spam and scams, content violations and copyright infringements.

LinkedIn Community Report: global actions taken on content that violated Professional Community Policies and User Agreement, January 2021 – December 2022

		2021 Jan-Jun	2021 Jul-Dec	2022 Jan-Jun	2022 Jul- Dec+	2023 Jan - Jun	2023 Jul - Dec
Global	Fake Accounts Stopped at registration	11.6m	11.9m	16.4m	44.7m	42.5m	46.3m
	Restricted proactively	3.7m	4.4m	5.4m	13.2m	15.1m	17.1m
	Restricted after report	85.7k	127k	190k	201k	196k	232.4k
	Content Violation Misinformation*	147.5k	207.5k	172.4k	138k	85.2k	53.8k

		2021 Jan-Jun	2021 Jul-Dec	2022 Jan-Jun	2022 Jul-Dec	2023 Jan - Jun	2023 Jul - Dec
Australia#	Fake Accounts Stopped at registration	54,883	45,983	81,533	149,591	112,767	253,569
	Restricted proactively	64,642	39,179	63,317	112,809	116,613	126,502
	Restricted after report	1,281	1,448	1,755	2,023	2,168	2,633
	Content Violation Misinformation*	2,149	6,007	3,946	1,656	969	571
	Misinformation content		219	151	79	13	17

	removals that were appealed by the content author						
	The number of appeals that were granted		3	3	3	1	3

+ Since July – December 2022, LinkedIn stopped more fake accounts compared to previous periods. Because of our [multidimensional approach](#) to combating fake accounts, the manner in which we catch fake accounts changed a bit from July 2022 onwards. With the rise of fraudulent activity taking place across the internet, LinkedIn continue to treat fake accounts as a top priority and invest in additional [verification features](#), [safety tools](#), and automated defenses to support safe experiences.

*Misinformation not reported as a separate category prior to 2020. Other content violation categories reported are harassment or abusive, adult, hateful or derogatory, violent or graphic, child exploitation.

#Australian data not reported separately prior to 2021.

Outcome 1b: Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.

Users can find information about the types of behaviours and content that will be prohibited and/or managed as follows:

- **Microsoft Advertising:** [Microsoft Advertising policies](#)
- **Microsoft Start:** [Microsoft Services Agreement, Community Guidelines](#)
- **LinkedIn:** [User Agreement, Professional Community Policies](#)

Outcome 1c: Users can report content and behaviours to Signatories that violate their policies under 5.10 through publicly available and accessible reporting tools.

In addition to the guidelines contained within the respective user agreements, **Bing Search**, **Microsoft Start**, **Microsoft Advertising** and **LinkedIn** have reporting mechanisms where users are able to flag problematic content.

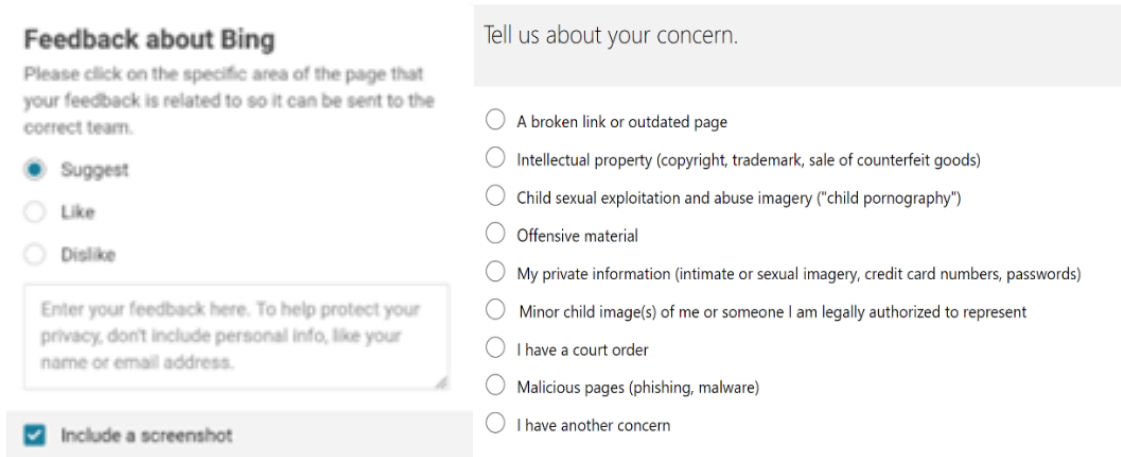
(O.1c) Bing Search

Bing Search has updated its "[Report a Concern](#)" and "Feedback" tools to include enhanced reporting for generative AI features as well as traditional web search. Bing Search's Report a Concern Form permits users to report third-party websites for a variety of reasons including disclosure of private information, spam and malicious pages, and illegal materials. Bing Search's "Feedback" tool, which is accessible on the lower right corner on a search results page, allows users to provide feedback on search results (including a screenshot of the results page) to Bing Search.

These tools have also been updated to make it easy for users to report problematic content they encounter while using Copilot in Bing by including the same "Feedback" button with

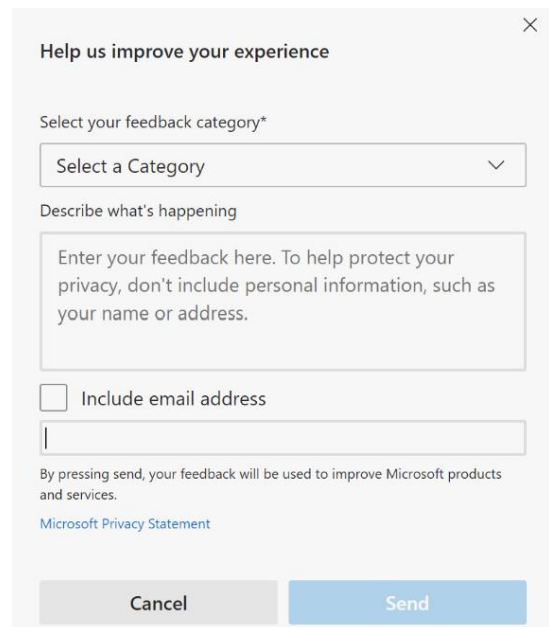
direct links to the respective service’s “Report a Concern” tool on the footer of each page of Copilot.

Depending on the nature of the feedback, Bing Search may take appropriate action, such as to engage in algorithmic interventions to ensure high authority content appears above low authority content in search results, remove links that violate local law or Bing policies, add answers, warnings or other media literacy interventions on certain topics, or remove auto-suggest terms.



(O.1c) Microsoft Start

Microsoft Start includes a feedback feature at the bottom of all pages (landing page and each article, see below), with Content Quality as one of the options in the drop-down menu. This feedback feature is also included in the Settings menu. In addition, each article includes a ‘Report an issue’ option, with ‘misleading title’, ‘outdated article’ and ‘suspected AI/bot created’ available as types of issues.



(O.1c) Microsoft Advertising

Microsoft Advertising enables users to report ads which may be in violation of its policies (e.g., ads that may contain malvertising, disallowed content, relevancy concerns, or sensitive content) through its [low quality ad submission and escalation form](#) (as shown below). Users of Bing Search and Microsoft Start can also report ads via the respective feedback functions on those services.

Low quality ad submission & escalation

Have you found an occurrence of a low quality ad on Microsoft Bing? Let us know! A low quality ad is one where the ad contains one or more of the following attributes:

- **Malvertising**: Describes advertising practices that have malicious intent to cause harm or defraud a user.
- **Disallowed content**: Refers to issues with landing page content/products/services that are not allowed in ads.
- **Relevancy concerns**: Poor relevancy can occur when an advertiser associates a keyword to a landing page or ad copy where no logical association exists (for example, a query for "Facebook" yields ad copy and a landing page for golf supplies).

Fill out the form below to submit an ad quality escalation.

*Required

Please enter your search query term *

Please enter the ad link (found on the Bing results page) *

This is not the display URL found in the ad. To copy the link:

1. Right click the ad title
2. Select Copy shortcut
3. Paste into the box below

Email *

Confirm email address *

Country/Region *

Ad attributes or issues

Please check the relevance, content or malvertising issues that are relevant to the ad(s) being escalated:

Disallowed content

- The ad's landing page has disallowed content The ad's landing page is promoting disallowed products or services
 Other disallowed content issue (explain in Comments section below)

Relevance

- The ad is not relevant to what I was looking for The landing page is not relevant to what I was looking for
 Ad copy does not make sense The display URL I saw in the ad does not match the landing page
 Other relevance issue (explain in Comments section below)

Page or site quality

- High percentage ads or links on the landing page Low value, sparse or limited content across the site
 This site redirects me to a completely unrelated location/domain
 Other page or site quality issue (explain in Comments section below)

Personally identifiable information (PII)

- Site asks me for personal information that I wouldn't expect to have to share Phishing

Malicious

- This site gave me a virus, or seems to host malware or spyware This site/business seems deceptive or fraudulent
 Other malicious issue (explain in Comments section below)

Landing page navigation

- Site changes browser preferences without my consent
 Site spawns multiple pop ups or pop ups that prevent me from leaving the site Landing page does not load
 I am getting a 'product not available' message
 Other landing page navigation issue (explain in Comments section below)

Sensitive content

- Ad exploits a sensitive tragedy, disaster, death or high profile news event, or is considered inappropriate given current events

Comments

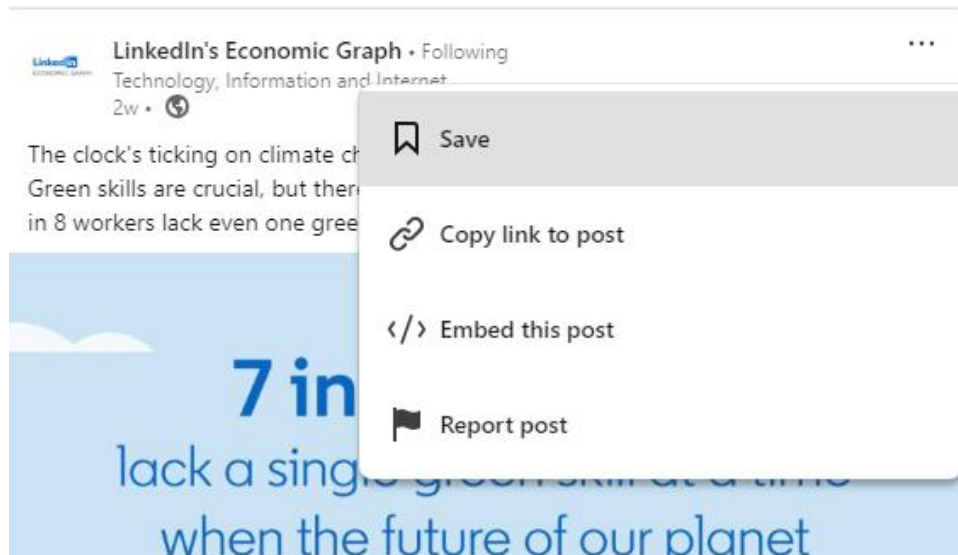
- I would like information, tips and offers about Microsoft Advertising. [Privacy Statement](#)

 I'm not a robot 

Submit

(O.1c) LinkedIn

If **LinkedIn's** members locate content they believe violates our Professional Community Policies, we encourage them to report it using the in-product reporting mechanism represented by the three dots in the upper right hand corner of a post on LinkedIn:



Report this post



Select an action



Provide feedback to change your feed

If you think this is inappropriate, you can give us feedback instead of reporting.



Report content for review

Tell us how this goes against our policies or request help for someone.



Misinformation is specifically called out as one of the reporting options.

Report this post ✕

Select a reason that applies

Harassment + Fraud or scam + Spam + Misinformation +

Hateful speech + Threats or violence + Self-harm + Graphic content +

Dangerous or extremist organizations + Sexual content + Fake account +

Child exploitation + Illegal goods and services + Infringement +

Next

Report this post ✕

Select a reason that applies

Harassment + Fraud or scam + Spam + **Misinformation ✓**

Hateful speech + Threats or violence + Self-harm + Graphic content +

Dangerous or extremist organizations + Sexual content + Fake account +

Child exploitation + Illegal goods and services + Infringement +

Next

Report this post ✕

You've selected the following reason

Misinformation
False content or information, including news stories, that present untrue facts or events as though they are true or likely to be true

Back Submit report

Reported content is generally reviewed by trained content reviewers. In addition, LinkedIn uses automation to flag potential violations including disclosure of private information, spam and malicious pages, and illegal materials content to our content moderation teams. If reported or flagged content is found to violate the Professional Community Policies, it will be removed from the platform.

When members use the above reporting process and choose to receive updates, LinkedIn communicates by email with the reporting member to confirm receipt of reports and provide updates about subsequent decisions. Members also generally receive notice in the event their content is removed from LinkedIn.

If members wish to appeal LinkedIn's decisions, they can request a second review and provide the reasons they believe LinkedIn's decision was incorrect. To begin that appeal process, members can log into their account and follow the onscreen messaging or reply to the message they received notifying them of the content removal.

Outcome 1d: Users will be able to access general information about Signatories actions in response to reports made under 5.11.

(O.1d) Bing Search, Microsoft Start, Microsoft Advertising

In addition to the sources detailed below, Microsoft regularly publishes information about the detection and removal of content that violates our policies or is subject to removal under local legal obligations in the [Digital Trust section of our Reports Hub](#).

(O.1d) LinkedIn

As noted in our response to O.1a above, the **LinkedIn [Community Report](#)** describes actions we take on content that violates our Professional Community Policies and User Agreement. It is published twice per year and covers the global detection of fake accounts, spam and scams, content violations and copyright infringements.

Outcome 1e: Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.

(O.1e) LinkedIn

LinkedIn has published several articles to explain to users how our recommender systems work, including:

- [Mythbusting the Feed: How the Algorithm Works](#)
- [Mythbusting the Feed: Helping our members better understand LinkedIn](#)
- [Keeping your feed relevant and productive LinkedIn Safety Series: Using AI to Protect Member Data](#)
- [Guide: Features to Help You Control Your Feed and Conversations](#)
- [Our approach to building transparent and explainable AI systems](#)



LinkedIn also makes easily accessible in the footer of every LinkedIn page a link out for "Recommendation Transparency", which links to a Help Centre article about how LinkedIn ranks content for the member. That article links to [more information](#) about how members can customize their feed, including the ability for members to sort their most relevant posts chronologically in desktop view.

Additionally, LinkedIn addresses automated processing and relevancy in the LinkedIn [User Agreement](#) at the end of Section 3.6 and in our [Help Centre](#).

Objective 2: Disrupt advertising and monetisation incentives for Disinformation.

Outcome 2: Advertising and/or monetisation incentives for Disinformation are reduced.

Microsoft strives to provide our customers with a positive online experience free from deceptive advertisements. Demonetisation is one of Microsoft's core Information Integrity [Principles](#), which outlines how we will not willfully profit from foreign cyber influence content or actors. Microsoft is working across our services to achieve this goal through policies and enforcement processes aimed at ensuring that the advertising and content served is clear, truthful, and accurate.

(O.2) Microsoft Advertising

In December 2022, **Microsoft Advertising** rolled out revised network-wide policies to avoid the publishing and carriage of harmful disinformation and the placement of advertising next to disinformation content. Such policies prohibit ads or sites that contain or lead to disinformation. To enforce this policy, we may use a combination of internal signals and trusted third-party data or information sources to reject, block, or take down ads or sites that contain disinformation or send traffic to pages containing disinformation. We may block at the domain level landing pages or sites that violate this policy. Please see our [main policy page](#).

Microsoft Advertising assesses the impact of its actions by reporting on the individual ads that we prevented from monetizing on web properties participating in the Microsoft Advertising network (i.e., "publisher sites" that use the Microsoft Advertising services to display ads on their properties), and the number web domains that we blocked from participating in our ad network.

Since the last reporting period, we have made additional system upgrades to further prevent ad calls on web domains that we blocked and eliminated impressions on these domains, thus enforcing our policies more effectively. Microsoft Advertising works with selected, trustworthy publishing partners and requires these partners to abide by strict brand safety-



oriented policies to avoid providing revenue streams to websites engaging in misleading, deceptive, harmful, or insensitive behaviours.

Microsoft Advertising's policies with respect to these publishers include a comprehensive list of prohibited content that ads cannot serve against. Prohibited content includes, but is not limited to:

- disinformation;
- sensitive content (e.g., extreme, aggressive, or misleading interpretations of news, events, or individuals);
- unmoderated user-generated content; and
- unsavoury content (such as content disparaging individuals or organisations).

Publishers are required to maintain a list of prohibited terms and provide us with information on their content management practices where applicable. In addition to content requirements, publishers are required to abide by restrictions against engaging in business practices that are harmful to users (e.g., distributing malware).

Advertisers who willingly or repeatedly violate our terms or policies are suspended from accessing the service and cannot service ads until they redress the violation.

(O.2) LinkedIn

LinkedIn prohibits misinformation and disinformation on its platform, whether in the form of organic content or in the form of advertising content.

LinkedIn's Professional Community Policies, which apply to all content on LinkedIn's platform expressly prohibit false and misleading content, including misinformation and disinformation. LinkedIn provides additional specific examples of false and misleading content that violates its policy via a Help Center article on [False or Misleading Content](#).

LinkedIn's [Advertising Policies](#) incorporate the above provision, and similarly prohibit misinformation and disinformation. In addition, LinkedIn's Advertising Policies also prohibit fraudulent and deceptive ads and require any claims made in an ad have factual support.

Of note, LinkedIn does not allow members to monetise or run ads against their content, nor does it offer an ad revenue share program. Thus, members publishing disinformation on LinkedIn are not able to monetise that disinformation or collect advertising revenue via LinkedIn.

LinkedIn members may also report ads that they believe violate LinkedIn's advertising policies and, when members report ads, LinkedIn's Advertising Review team reviews them. To report an ad, members can click on the three-dot icon in the upper right-hand corner of every ad and select the "Hide or report this ad" option.

LinkedIn provides a range of information and tools to give advertisers transparency and control regarding the placement of their advertising. For example, for ads on the LinkedIn platform, LinkedIn publishes a Feed Brand Safety score for advertisers and the public. The Feed Brand Safety score measures the number of ad impressions on the LinkedIn platform that appeared adjacent to – that is, immediately above or below within the LinkedIn feed – content removed for violating LinkedIn’s Professional Community Policies, including disinformation. From July through December 2023, the Feed Brand Safety score was 99%+ safe. More information about [LinkedIn’s Feed Brand Safety Score](#).

Objective 3: Work to ensure the security and integrity of services and products delivered by Digital platforms.

Outcome 3: The risk that inauthentic user behaviours undermine the integrity and security of services and products is reduced.

In addition to the actions detailed in Objective 1 (Outcomes 1a, 1b and 1c), **Bing Search**, **Microsoft Advertising**, and **LinkedIn** reduce the risk of inauthentic user behaviours through the measures detailed below.

(O.3) Bing Search

The “Abuse and Examples of Things to Avoid” section of the [Bing Webmaster Guidelines](#) details the policies intended to maintain the integrity of Bing Search. Bing’s general spam policies prohibit certain practices intended to manipulate or deceive the Bing search algorithms.

Bing may take action on websites employing spam tactics or that otherwise violate the Webmaster Guidelines, including by applying ranking penalties (such as demoting a website or delisting a website from the index). However, it is important to clarify that in search it is not feasible to distinguish between spam tactics employed by malicious actors specifically for the purpose of spreading disinformation and other types of spam.

In addition to enforcing its spam policies, Bing takes actions to promote high authority, high quality content and thereby reduce the impact of disinformation appearing in Bing search results. Among other initiatives, this includes:

- continued improvement of its ranking algorithms to ensure that the most authoritative, relevant content is returned at the top of search results;
- regular review and actioning of disinformation threat intelligence;
- contributing to and supporting the research community; and
- implementation and enforcement of clear policies concerning the use of manipulative tactics on Bing Search.



Although the Bing search algorithms endeavour to prioritise relevance, quality, and credibility in all scenarios, in some cases Bing identifies a threat that undermines the efficacy of its algorithms. When this happens, Bing employs “defensive search” strategies and interventions to counteract threats in accordance with its trustworthy search principles to help protect Bing users from being misled by untrustworthy search results and/or inadvertently being exposed to unexpected harmful or offensive content.

In addition to defensive search, Bing Search regularly monitors for violations of its Webmaster Guidelines, including attempts to manipulate the Bing search algorithms through prohibited practices such as cloaking, link spamming, keyword stuffing, and phishing.

The above measures also support Copilot in Bing. Responses provided by the Copilot feature are “grounded” in search results, which are based on the same ranking algorithms and moderation infrastructure that are used by Bing’s traditional web search, and, as such, benefit from Bing’s longstanding safety infrastructure described above. Nonetheless, Microsoft recognizes that generative AI technology may also raise new risks and possibilities of harm that are not present in traditional web search and has supplemented its existing threat identification and mitigation processes with additional risk assessments and mitigation processes based on [Microsoft’s Responsible AI](#) program.

(O.3) Microsoft Advertising

Microsoft Advertising employs a robust filtration system to detect bot traffic.

- This system uses various algorithms to automatically detect and neutralise invalid or malicious online traffic which may arise from or result in click fraud, phishing, malware, or account compromise.
- The system is supported by several teams of security engineers, support agents, and traffic quality professionals who continually develop and improve monitoring and filtration.
- Support teams work closely with advertisers to review complaints around suspicious online activity and across internal teams to verify data accuracy and integrity.

(O.3) LinkedIn

LinkedIn’s professional focus shapes the type of content we see on platform. People tend to say things differently when their colleagues and employer are watching. Accordingly, our members do not tend to use LinkedIn to engage in the mass dissemination of misinformation, and bad actors generally need to create fake accounts to peddle misinformation.

To ensure their content reaches a large audience, bad actors need to either connect with real members or post content that real members will like— both of which are hard to achieve on LinkedIn given our professional focus. The mass dissemination of false information, as well as artificial traffic and engagement, therefore, requires the mass creation of fake accounts, which we have various defences to prevent and limit.

To evolve to the ever-changing threat landscape, our team continually invests in new technologies for combating inauthentic behaviour on the platform. We are investing in artificial intelligence technologies such as advanced network algorithms that detect communities of fake accounts through similarities in their content and behaviour, computer vision and natural language processing algorithms for detecting AI-generated elements in fake profiles, anomaly detection of risky behaviours, and deep learning models for detecting sequences of activity that are associated with abusive automation.

LinkedIn acts vigilantly to maintain the integrity of all accounts and to ward off bot and false account activity (including “deep fakes”).

LinkedIn enforces the policies in its [User Agreement](#) prohibiting the use of “bots or other automated methods to access the Services, add or download contacts, send or redirect messages” through:

- having a dedicated Anti-Abuse team to create the tools to enforce this prohibition;
- using automated systems detect and block automated activity;
- imposing hard limits on certain categories of activity commonly engaged in by bad actors;
- detecting whether members have installed known prohibited automation software;
- conducting manual investigation and restriction of accounts engaged in automated activity;
- partnering with the broader Microsoft organisation to develop technological solutions for protecting content provenance and identification of deep fakes;
- investing in and using AI to detect coordinated inauthentic activity and communities of fake accounts through similarities in their content and behaviour;
- using third party fact checking sites during the human content review process when suspected deepfakes are flagged or found on the platform; and
- “hashing” known instances of deepfake content, which can be used to find copies of the same content on our platform.



Objective 4: Empower consumers to make better informed choices of digital content.

Outcome 4: Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation.

Microsoft is committed to helping our users make informed decisions about content. This includes providing our customers with tools to help them evaluate the trustworthiness of that content.

Microsoft is working both internally and with third parties to provide new tools and implement new technologies across our services to assist our customers in identifying trustworthy, relevant, authentic, and diverse content, including in news, search results, and user-generated material.

(O.4) Bing Search

Bing Search offers a number of tools to help users understand the context and trustworthiness of search results. Even in circumstances where a user is expressly seeking low authority content (or if there is a data void so little to no high authority content exists for a query), Bing Search provides tools to users that can help improve their digital literacy and avoid harms resulting from engaging with misleading or inaccurate content.

Bing Search is enhanced with various features to help users navigate complex information environments with confidence, for example, in addition to what we noted in last year's report:

- Bing Search Intelligent Answers also provides users with informative panels and direct answers to certain search queries, and is now available in 100 languages.
- Bing Search's "Knowledge Cards" feature gives users a single view of authoritative information on a specific topic and are typically displayed at the top of the SERP page.
- Bing Search's [Page Insights](#) feature also helps provide users with information and context about websites contained in the search results. The feature, which appears as a light bulb image next to certain search results, provides users with additional information about the site and its contents from third party information sites such as Wikipedia
- Bing Search ingests tags for fact-check articles using the ClaimReview open schema to help users find fact checking information and warns users with red "flags" when fact-checked claims or content appearing in search results has been determined to be false or unfounded by third-party fact checkers;



- Microsoft also partners with NewsGuard to help users evaluate the quality of the news they encounter online. NewsGuard launched in Australia in March 2023 and is available as a free plug-in for the Microsoft Edge web browser (it is also available for other browsers including Chrome and Firefox), and users of the Edge mobile application on both iOS and Android can enable NewsGuard ratings in their app settings. NewsGuard [reported](#) rating the news and information sites that account for 92% of engagement with the news in Australia and New Zealand. For users with the NewsGuard plug-in, Bing Search results include NewsGuard Reliability ratings that lead to a pop-up screen with more site information;
- Microsoft continues to offer Search Coach as a free app in Microsoft Teams to help educators and students form effective queries and identify reliable resources. It is designed to teach information literacy skills in a safe, secure, and ad-free environment.

Over the reporting period Microsoft has:

- Strengthened its partnerships with third-party organisations, including the News Literacy Project and The Trust Project, to fund media literacy campaigns while continuing introductory calls with new organizations to grow additional campaigns' reach to new markets;
- Provided pro-bono advertising space across Microsoft surfaces to disseminate the literacy campaigns and helped garner millions of impressions per month;
- Helped educators build AI literacy and make the most of AI capabilities, we introduced a free module on Microsoft Learn: [Enhancing teaching and learning with Copilot](#). This module is designed to guide educators through available features, learn how to create and iterate on prompts, and use expertise to evaluate responses for quality and credibility; and
- Supported the creation of [The Investigators](#), a new world for Minecraft Education that helps students build information and media literacy through game-based learning. Launched in English, the game and support materials are being localized to become available globally in 28 languages to millions of students and teachers in 2024.

Copilot in Bing

In addition to the features available for core search experiences, Copilot in Bing also provides information to help educate users on the uses and limitations of generative AI-driven search experiences, such as by reminding users that they are interacting with a generative-AI system and that mistakes can occur (see below):

Copilot uses AI. Check for mistakes. [Legal Terms](#) | [Privacy and Cookies](#) | [FAQ](#)

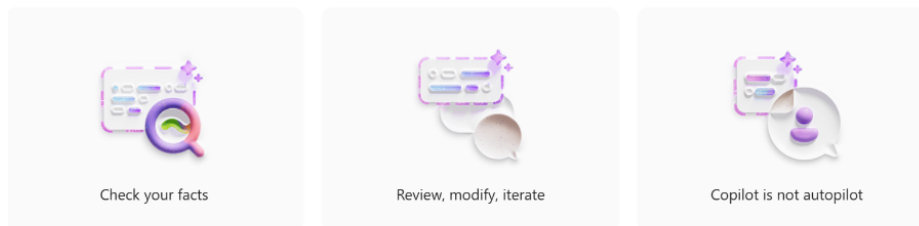
[The Copilot in Bing FAQs](#) and similar explanatory documents like [blog posts](#) and [Copilot in Bing: Our Approach to Responsible AI](#) also help educate users on the nature of AI-driven search experiences and the uses, safeguards, and limitations of this emerging technology.

For example, [the Copilot FAQ answer](#) to “Are Copilot’s AI-generated responses always factual?” explains: “Copilot aims to base all its responses on reliable sources - but AI can make mistakes, and third-party content on the internet may not always be accurate or reliable. Copilot will sometimes misrepresent the information it finds, and you may see responses that sound convincing but are incomplete, inaccurate, or inappropriate. Use your own judgment and double check the facts before making decisions or taking action based on Copilot’s responses.”

In another prominent example, another [Copilot support page](#) reminds users prominently that while they “lead the way”, they need to check facts, review, and avoid simply relying on AI as an “autopilot” (shown below). Additionally, Microsoft has released a [Classroom Toolkit](#) for teachers that encourages responsible education around generative AI tools including the importance of fact checking.

You lead the way

Unleash your creativity and get things done with Copilot by your side. Since AI-generated content may be incorrect, here are a few things to remember...



Microsoft also offers meaningful resources for users interested in learning more about generative AI features and tools, including Copilot, through blog posts, articles, information hubs, and support pages. In addition to teaching AI basics and how-tos, these resources reiterate the importance of checking AI-generated materials and understanding the strengths and limitations of AI. See e.g., [Microsoft AI help & learning](#). For example, a recent Copilot support article ([Unleash your productivity with AI and Microsoft Copilot – Microsoft Support](#)) has a section specifically directing users to “Be Aware of AI limitations”, which explains to users that AI-produced content and outputs may contain inaccuracies, “biases, or sensitive materials because they were trained on information from the internet, as well as other sources. AI may not know about recent events yet, and struggles to understand and interpret sarcasm, irony, or humour.”

Microsoft is committed to providing resources, educational materials, and guides so that users can develop literacy when interacting with AI systems and will continue to explore ways to further educate the public on important generative AI topics.



(O.4) Microsoft Start

Microsoft Start clearly labels the sources of news articles and distinguishes advertising to enable users to readily differentiate this from other content.

(O.4) LinkedIn

As the world around us changes, **LinkedIn** continues to evolve and adapt our systems and practices for combating misinformation and other inauthentic behaviour on our platform, including to respond to the unique challenges presented by world events.

In addition to broader measures, LinkedIn has taken steps to tackle disinformation in connection with unfolding world events. LinkedIn's in-house editorial team provides members with trustworthy content regarding global events. LinkedIn does not prioritise any news sources in our feed, but in crisis situations, we will use search banners to point members to reputable sources of information.

As mentioned above, Microsoft has partnered with NewsGuard to provide a free plug-in for the Microsoft Edge web browser (also available for other browsers). LinkedIn members are also able to benefit from NewsGuard via this plug in which enables LinkedIn members to benefit from NewsGuard's reliability rating, where available, when browsing news posts from news and information sites rated by NewsGuard.

Further, in October 2022, LinkedIn began [offering](#) an "About this profile" feature that shows users when a profile was created and last updated, along with whether the member has verified a phone number and/or work email associated with their account. Over the past year, LinkedIn also has been rolling out a range of [free verifications](#), which allow our members to verify certain information about themselves, like their association with a particular company or educational institution or their identity (through one of LinkedIn's verification partners).

The above features can be strong user empowerment tools. Specifically, they can provide our members valuable authenticity signals to help them make more informed decisions about what content and individuals they engage with online.

(O.4) Other contributions and measures

Globally, Microsoft also has a number of programs to proactively combat disinformation on our services and empower users.

Microsoft's commitments and actions under the Tech Accord

Microsoft and LinkedIn are two of 20 companies that announced a new [Tech Accord to Combat Deceptive Use of AI in 2024 Elections](#). Microsoft has already [taken steps to meet the commitments in the Tech Accord](#) by further implementing content provenance, establishment of reporting channels and improved detection capability. For example:

- Microsoft is harnessing the data science and technical capabilities of our AI for Good Lab and Microsoft Threat Analysis Center teams to better detect deepfakes on the internet. We will call on the expertise of our Digital Crimes Unit to invest in new threat intelligence work to pursue the early detection of AI-powered criminal activity.
- In addition, Microsoft will launch [Content Credentials as a Service](#) to enable political candidates around the world to digitally sign and authenticate media using the Coalition for Content Provenance and Authenticity's (C2PA) digital watermarking credentials.

We combined this work with the launch of an expanded Digital Safety Unit. This will extend the work of our existing digital safety team, which has long addressed abusive online content and conduct that impacts child or that promotes extremist violence, among other categories. This team has special ability in responding on a 24/7 basis to weaponized content from mass shootings that we act immediately to remove from our services. The accord's commitments oblige Microsoft and the tech sector to continue to engage with a diverse set of global civil society organizations, academics, and other subject matter experts. These groups and individuals play an indispensable role in the promotion and protection of the world's democracies.

Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.

Outcome 5: Users are better informed about the source of Political Advertising.

(O.5) Microsoft Advertising

Under our Advertising Policies, **Microsoft Advertising** prohibits political advertising. This includes ads for election-related content, political candidates, parties, ballot measures, and political fundraising globally; similarly, ads aimed at fundraising for political candidates, parties, political action committees (PACs), and ballot measures also are barred.



All Microsoft and third-party services that rely on Microsoft Advertising to serve advertisements on their platforms benefit from these robust, and robustly enforced, set of policies.

Specifically, Microsoft Advertising employs dedicated operational support and engineering resources to enforce restrictions on political advertising using a combination of proactive and reactive mechanisms.

- On the proactive side, Microsoft Advertising has implemented several processes designed to block political ads from showing across its advertising network, including restrictions on certain terms and from certain domains.
- On the reactive side, if Microsoft Advertising becomes aware that an ad suspected of violating its policies is being served to our publishers—for instance, because someone has flagged that ad to our customer support team—the offending ad is promptly reviewed and, if it violates our policies, taken down.

Microsoft Advertising’s policies also prohibit certain types of advertisements that might be considered issue based. More specifically, “advertising that exploits political agendas, sensitive political issues or uses ‘hot button’ political issues or names of prominent politicians is not allowed regardless of whether the advertiser has a political agenda,” and “advertising that exploits sensitive political or religious issues for commercial gain or promote extreme political or extreme religious agendas or any known associations with hate, criminal or terrorist activities” are also prohibited.

(O.5) LinkedIn

LinkedIn does not accept political advertising. LinkedIn’s Advertising Policies globally prohibit political ads which:

- advocate for or against a particular candidate, party or ballot proposition or are otherwise intended to influence an election outcome;
- fundraise for or by political candidates, parties, ballot propositions or PACs or similar organisations; and
- exploit a sensitive political issue even if the advertiser has no explicit political agenda.

All ads are subject to review for adherence to policy before being approved to run. LinkedIn has also introduced features making it simple for members to [report advertisements](#) that violate LinkedIn’s policies; LinkedIn reviews such reports and removes offending advertisements from its platform.

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.

Outcome 6: Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation.

A non-exhaustive list of Microsoft’s ongoing collaborations with the broader research community in this space include:

Bing Search ORCAS dataset	<p>Bing Search provides researchers with access to ORCAS: Open Resource for Click Analysis in Search a click-based dataset associated with the TREC Deep Learning Track, which provides 18 million connections to 10 million distinct queries and is available to researchers.</p>
<p>Responsible AI Toolbox</p>	<p>As a leader in research in Responsible AI, Microsoft provides a range of tools and resources dedicated to promoting responsible usage of artificial intelligence to allow practitioners and researchers to maximize the benefits of AI systems while mitigating harms. For example, as part of its Responsible AI Toolbox, Microsoft provides a Responsible AI Mitigations Library, which enables practitioners to more easily experiment with different techniques for addressing failure (which could include inaccurate outputs), and the Responsible AI Tracker, which uses visualizations to show the effectiveness of the different techniques for more informed decision-making. These tools are available to the public and research community for free.</p>
<p>Partnership on AI</p>	<p>Microsoft is a partner in Partnership on AI which works to better understand and address the emerging threat posed by the use of AI tools to develop malicious synthetic media (i.e., deep fakes).</p>
<p>MS MARCO</p>	<p>Bing Search makes information available to the research community to improve search results by making data sets like its MS MARCO publicly available. Bing Search provides researchers and the public with access to MS MARCO, a collection of datasets focused on deep learning in search that are derived from Bing queries and related data. Research organisations can gain access to the MS MARCO datasets instantaneously via the MS MARCO homepage. The MS MARCO dataset has been cited in over 1400 research papers since its release and has been used for a range of research issues, including in relation to misinformation and disinformation. Because the dataset is provided open source, the extent to which it has been used for disinformation related research purposes cannot easily</p>

	<p>be ascertained. However, the dataset has been cited in various academic papers concerning misinformation and disinformation, including:</p> <ul style="list-style-type: none"> • “Retrieving Supporting Evidence for Generative Question Answering”, SIGIR-AP '23: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, November 2023. • “Cross-Genre Retrieval for Information Integrity: A COVID-19 Case Study”, In: Yang, X., <i>et al.</i> Advanced Data Mining and Applications. ADMA 2023. Lecture Notes in Computer Science), vol 14180. Springer, Cham, November 2023. • “Personas as a Way to Model Truthfulness in Language Models” New York University, ETH Zurich, et al. arXiv:2310.18168, October 2023.
MS MARCO Web Search	<p>Bing Search also recently released the MS MARCO Web Search dataset, a large-scale information-rich Web dataset, featuring millions of real clicked query-document labels. This dataset closely mimics real-world web document and query distribution, provides rich information for various kinds of downstream tasks. MS MARCO Web Search further contains 10 million unique queries from 93 languages with millions of relevant labeled query-document pairs collected from the search log of the Microsoft Bing search engine to serve as the query set.</p>
<p>Other publicly shared datasets</p>	<p>Bing Search also offers use of Bing APIs to the public, which include services such as Bing Image Search, Bing News Search, Bing Web Search. Bing Search provides free access to these APIs for up to 1,000 transactions per month, which may be leveraged by the research community.</p> <ul style="list-style-type: none"> • Given the open nature of the Bing Search index and public nature of search results, researchers can use Bing to run specific queries and analyse results (unlike social media which may require private accounts or connections between users to access certain materials).

Democracy Forward Initiative

Microsoft believes technology companies have a responsibility to help protect democratic processes and institutions globally. Though threats to democracy have always existed, the tactics of adversaries are constantly evolving. Microsoft is protecting open and secure



democratic processes by providing services and technology to secure critical institutions, protect electoral processes from cyberattacks, and build public trust in voting procedures.

Microsoft's Democracy Forward Initiative is an innovative effort to protect democratic institutions and processes from hacking, to explore technological solutions to protect electoral processes, and to defend against disinformation.

Microsoft's election principles

In November 2023, we [announced](#) a set of election key principles and several tangible steps to protect voters, candidates, political campaigns, and election authorities including:

- Launching Content Credentials as a Service that enables political campaign users to digitally sign and authenticate media with the Coalition for Content Provenance and Authenticity's ([C2PA](#)) content credentials digital watermark.
- Deploying a Campaign Success Team to support campaigns navigating the AI landscape, and Microsoft's M365 for Campaigns and AccountGuard services.
- Launching an Elections Communications Hub for elections officials across the globe to share security concerns on Microsoft's platforms, including potential incidents of disinformation.
- Launching a dedicated [Microsoft Elections](#) page where a political candidate can report to us a concern about a deepfake of themselves.

Microsoft's Democracy Forward team continues to expand its collaborations with organizations that provide information on authoritative sources, ensuring that queries about global events will surface reputable sites. Microsoft works with Reporters Without Borders (RSF) and their Journalism Trust Initiative (JTI) data to proactively promote trusted sources of news around the world.



Objective 7: Signatories publicise the measures they take to combat Disinformation and Misinformation.

Outcome 7: The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.

Our reporting under this code is available on the Microsoft Australia News Centre and on DIGI's website.

Microsoft also releases other information about our initiatives globally to combat disinformation:

(O.7) Bing Search, Microsoft Start, Microsoft Advertising, LinkedIn

Microsoft On the Issues	<p>Blog contains announcements on technology policy issues, including disinformation.</p> <p>For example, our response to the invasion of Ukraine, various elections around the world, video authenticator technology, release of digital trust reports, are all posted on the blog.</p>
Microsoft Reports Hub	<p>Transparency reports include Digital Safety Content Report and Government Requests for Content Removal Report.</p>
Microsoft Digital Defense Report	<p>Report encompasses learnings from security experts, practitioners, and defenders at Microsoft to empower people everywhere to defend against cyberthreats. Includes dedicated section on disinformation.</p>
Microsoft's Inaugural Responsible AI Transparency Report	<p>Provides insight into how we build applications that use generative AI; make decisions and oversee the deployment of those applications; and learn, evolve, and grow as a responsible AI community.</p>
LinkedIn Transparency Center	<p>Community Report</p> <p>Government Requests Report</p>
LinkedIn Blog	<p>Blog contains information on actions to combat disinformation, including New LinkedIn profile features help verify identity, detect and remove fake accounts, boost authenticity, How We're Protecting Members From Fake Profiles, Automated Fake Account Detection, and An Update on How We Keep Members Safe</p>

Conclusion

Microsoft is dedicated to contributing to a reliable information ecosystem by implementing our policies, advancing research and innovation in emerging technologies, and engaging in collaboration with our partners, the academic community, and our users. This report outlines the measures being taken by Bing Search, Microsoft Start, Microsoft Advertising, and LinkedIn to mitigate and interrupt the spread of disinformation and misinformation. It also highlights the company's endeavours to fulfill the objectives and pledges of the Australian Code of Practice on Disinformation and Misinformation.

Australian Code of Practice on Disinformation and Misinformation
Redbubble Inc.
Annual Transparency Report
1 January 2023 - 31 December 2023

Summary

Redbubble is a global artist marketplace dedicated to giving independent artists a meaningful way to sell their creations. Redbubble hosts user-generated content uploaded by artists and provides them with online tools to upload their art and to design and sell products printed with their art to their customers worldwide.

Redbubble recognises the harm that arises from the spread of misinformation and disinformation and is committed to preventing, detecting and removing such harmful information from its marketplace. In doing so, Redbubble strives to balance the fostering of artistic freedom with the goal of preventing the spread of disinformation and misinformation through its platform.

By deploying a global content moderation team as well as scalable technologies like duplicate detection, image matching, keyword and text-in-image detection, as well as machine learning and AI, Redbubble is well positioned to efficiently tackle problematic content on its platform.

This report provides transparency into the measures that Redbubble takes to prevent, detect and remove disinformation and misinformation on its marketplace.

Commitments under the Code

Outcome 1a: Reducing harm by adopting scalable measures	Opt In
Outcome 1b: Inform users about what content is targeted	Opt In
Outcome 1c: Users can easily report offending content	Opt In
Outcome 1d: Information about reported content available	Opt In
Outcome 1e: Information about recommender engines	Opt In

Objective 2: Disrupt advertising and monetisation incentives for disinformation	Opt In
Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms	Opt In
Objective 4: Empower consumers to make better informed choices of digital content.	Opt In
Objective 5: Improve public awareness of the source of political advertising carried on digital platforms	Opt Out
Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research	Opt In
Objective 7: Signatories will publicise the measures they take to combat Disinformation	Opt In

Reporting against commitments

Outcome 1a: Reducing harm by adopting scalable measures

Redbubble prohibits users from uploading harmful disinformation and misinformation to the marketplace in the Redbubble User Agreement, published at <https://www.redbubble.com/agreement>, and the Redbubble Community and Content Guidelines at: <https://help.redbubble.com/hc/en-us/articles/202270929>.

Every image uploaded and keyword generated on the Redbubble platform must pass through one or more scalable detection technologies. Continuous improvements to tooling allow Redbubble to more accurately and efficiently detect and remove harmful misinformation and disinformation from its marketplace. Current scalable technologies include:

- Duplicate detection identifies previously moderated images that users may try to re-upload
- Image matching detects content similar to images known for disinformation or misinformation
- Keyword detection in text-based user-generated fields, like titles, tags, and descriptions to catch keywords linked to disinformation and misinformation

- Text-in-image matching spots text-based misinformation and disinformation within images
- Machine learning pinpoints user accounts linked to networks known for violating Redbubble policies
- Artificial intelligence recognizes users' keyword tagging patterns associated with disinformation and misinformation

The Redbubble Content Safety Team proactively screens the marketplace on a daily basis for potential misinformation and disinformation detected by Redbubble's tools, and the team removes content that it determines violates our User Agreement and Community Guidelines.

When the Content Safety Team makes decisions relating to content that potentially includes misinformation or disinformation, the team considers Redbubble policies, past decisions (to ensure consistency of approach and decision-making) and further research particularly in relation to new or emerging topics. The Content Safety Team's framework for content review is built on clear criteria and a repeatable and scalable workflow. This allows the team to make moderation decisions in an unbiased and consistent manner. Furthermore, policies and decision-making frameworks undergo continuous review and refinement to ensure their ongoing effectiveness.

Redbubble proactively monitors and screens for over 350 different content safety topics on its platform, spanning issues like incitement of violence, racism, disinformation and misinformation.

The following are examples of misinformation and disinformation topics that Redbubble screens for:

- Medical misinformation, such as anti-vaccine propaganda that may encourage the spread of communicable disease;
- Denials of real-world catastrophes, such as the Holocaust; and
- Political misinformation, such as false political conspiracy theories that are linked to real-world harm.

The Redbubble Content Safety team makes use of credible and trusted news sources in determining the boundaries of Disinformation and Misinformation, including review of independent, non-partisan fact checking sites, including:

- *Medical Misinformation*: Redbubble consults reports from a range of leading global authorities that are guided by scientific research to advise on everything from 'Plandemic' misinformation to false claims about Hydroxychloroquine. Sources include:
 - US Food & Drug Administration (FDA): <https://www.fda.gov>
 - Centers for Disease Control & Prevention: (CDC): <https://www.cdc.gov>
 - World Health Organization (WHO): <https://www.who.int>

- *Harmful Political Misinformation:* Redbubble consults non-partisan research centers and independent nonprofits run by professional researchers to guide decisions on political messaging that may cause real world harm. Sources include:
 - [FactCheck.org](https://factcheck.org)
 - Snopes: <https://www.snopes.com>
 - Sunlight Foundation: <https://sunlightfoundation.com>

To create screening guidelines, the team uses the above sources to compile training content for content review teams and quality assurance of moderation decisions.

Content Removal Trends

We continue to focus our efforts on content that may violate our Community Guidelines on Harmful Misinformation. We define this as any misleading or false information that harms or significantly threatens public health and safety, or where the intent is to cause fear and suspicion about a topic that can cause real-world harm. In 2023, we detected and removed 849 designs uploaded to the platform that contained harmful misinformation.

The most notable trend over the past few years is the increase and subsequent drop in instances of medical misinformation being uploaded to the platform. There was a sharp increase of such content in 2021 during the COVID-19 pandemic, followed by a gradual decline in subsequent years.

Subject matters such as election integrity, harmful misinformation about voting procedures and uploads condoning and perpetuating election-related real-world violence may increase again in the 2024 U.S. election cycle.

Year	Number of designs moderated for containing harmful misinformation
2021	10,811
2022	3,961
2023	849

Outcome 1b: Inform users about what content is targeted

The communication to users of what constitutes misinformation and disinformation is important in stopping its spread. To this end, Redbubble publishes content rules in its Community Guidelines and in various Redbubble Help Centre articles. In the Redbubble User Agreement, users are required to adhere to the Community Guidelines and represent and warrant that the products they sell are free from misinformation and disinformation. The User Agreement states that Redbubble reserves the right to review and in its sole discretion remove any such content from the website and terminate user accounts.

The Community Guidelines detail Redbubble's rules regarding user behavior and content on its marketplace and are made public in the spirit of open communication with artists and their customers. The relevant prohibition under the Community Guidelines addresses misinformation and disinformation in the following terms:

“Harmful misinformation is not permitted. We define this as any misleading or false information that harms or significantly threatens public health and safety, or where the intent is to cause fear and suspicion about a topic that can cause real-world harm.”

The Community Guidelines are expressly noted to be adaptable and subject to refinement over time as the environment and circumstances change.

The Redbubble User Agreement provides further information to artists proposing to upload content to the marketplace. It provides that an artist uploading content represents and warrants that:

“The content does not contain material that is harmful, abusive, inflammatory or otherwise objectionable; and the content is not misleading and deceptive and does not offer or disseminate fraudulent schemes or promotions.”

Additional information for users is included in the Redbubble Help Centre's Community Guidelines FAQ located at: <https://help.redbubble.com/hc/en-us/sections/4404750122004-Community-Guidelines-FAQ>. The Content Safety team adds to this FAQ on a regular basis to provide helpful information in response to commonly asked questions regarding disinformation and misinformation and other content safety topics.

Outcome 1c: Users can easily report offending content

Redbubble requests that all users of our marketplace flag behavior or content that contravenes the Community Guidelines through one of the reporting functions on our site. The User Agreement prompts users to report such content:

“Please help us by letting us know straight away about any inappropriate Content you see on the Marketplace. You can do this by clicking the "Inappropriate Content" link displayed on each Product listing page.”

Redbubble provides a simple and accessible reporting tool for users. On every product listing page created by sellers on the marketplace, a prominent “Report Content” link is provided.

This link directs users to a web form (shown below) where they can quickly and easily report content that they believe falls outside of Redbubble’s policies, including works that may contain disinformation or misinformation.

Inappropriate Content

Note: **Your information will *not* be relayed to the author**, you will remain anonymous.

If you'd like to report [redacted] by [redacted] as inappropriate, please complete this form and we'll make sure the review team takes a look.

Why are you reporting this content as inappropriate?
Please Select... ▾

If the content does not meet these guidelines, please provide any additional comments or information and click "send".

SEND

How does this work?

When you report a concern a notification is sent to the Redbubble objections team. We review the content and follow up in cases where the content falls outside Redbubble's guidelines. Due to the volume of emails the team receives, we cannot respond to every query regarding these reports but please rest assured we do check every single report carefully and we'll be in touch if we need any further information. If you would like more insight into the guidelines we apply, further information is available in our [community guidelines](#). Thanks again!

The Content Safety team regularly reviews all user reports to assess whether the content or uploading account should be subject to moderation or other actions. This reporting tool continues to be an important way to detect content that violates Redbubble policies, including disinformation and misinformation.

Outcome 1d: Information about reported content available

In 2023, users submitted over 15,000 reports using the reporting functionality described in Outcome 1c, which includes but is not limited to reports of misinformation or disinformation.

Outcome 1e: Information about recommender engines

Product recommendations to users of the Redbubble marketplace are primarily based on keyword-matching algorithms that connect the text-based title, tags and description generated by the artists with user search and navigation behavior, such as prior keywords the

user searched for and keyword similarities between content they clicked on and additional content that may interest them.

Objective 2: Disrupt advertising and monetisation incentives for disinformation.

Redbubble reduces monetisation incentives from artist sales of products on the marketplace by swiftly detecting and removing accounts and content that violates Redbubble policies. These measures are described in more detail throughout this report. If a user violates the Redbubble User Agreement or Community Guidelines, their violative content will be removed and their accounts will be subject to account penalties, up to and including account termination. Networks of connected accounts will also be terminated.

Redbubble also disrupts incentives by using keyword blocking tools that prevent content tagged with terms related to misinformation and disinformation from appearing on offsite marketing platforms where artists promote their products and generate sales. This blocklist covers thousands of keywords related to content safety topics, including hundreds of terms related to misinformation and disinformation.

Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms.

At the account-level, Redbubble uses third-party account abuse detection software that combines machine learning with a global network of data to detect users who are likely to violate the Redbubble User Agreement. This software uses data points that are customized to the Redbubble marketplace and allow room for adjustments based on emergent trends. This is an effective tool for maintaining content integrity on Redbubble and detecting users who are likely to upload content that perpetuates disinformation or misinformation. For example, this tool has helped detect users who use bots to create networks of multiple accounts and attempt to upload large amounts of images intending to sell products related to trending topics that may cause public harm. In 2023, over 350,000 accounts were blocked or removed by these measures.

Objective 4: Empower consumers to make better informed choices of digital content.

The measures and tools discussed throughout this report summarize the ways that Redbubble detects and removes misinformation and disinformation, which mitigates the risk of harmful content and accounts that consumers could be exposed to on its marketplace.

Redbubble also puts artists front and center in its public communications, such as web copy and promotional materials, to ensure that consumers understand that images have been uploaded by independent artists and that the products offered on the marketplace are designed and sold by artists.

The Redbubble User Agreement makes it clear to users that artists are responsible for the content they upload, the products they offer for sale in their shops, and the titles, tags and descriptions they write to describe their products.

Email marketing materials contain the following statement: “All products on the Redbubble marketplace are designed and sold by independent artists”, and every product listing page created by artists contains the words “designed and sold by [artist’s username]”.

Objective 5: Improve public awareness of the source of political advertising carried on digital platforms.

Redbubble has opted out of Objective 5, because political advertising is not considered to apply to the Redbubble business, and Redbubble does not sell ad space to parties conducting political advertising.

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.

Redbubble is open to supporting independent research that has the purpose of improving public understanding of disinformation and misinformation. At this time, Redbubble does not provide financial support to third-party strategic research in this area.

Objective 7: Signatories will publicise the measures they take to combat Disinformation.

Redbubble will continue to publish these transparency reports, which are accessible to the public on the DIGI website at digi.org.au/disinformation.



Australian Code of Practice on Disinformation and Misinformation

TikTok

Annual Transparency Report

January 2023 – December 2023



Table of Contents

<i>Summary</i>	2
<i>TikTok's Commitments under the Code</i>	5
<i>Reporting against 2023 commitments</i>	7
Objective 1: Provide safeguards against Harms that may arise from Disinformation and Misinformation.....	7
Outcome 1a: Reducing harm by adopting scalable measures	7
Outcome 1b: Inform users about what content is targeted	10
Outcome 1c: Users can easily report offending content.....	12
Outcome 1d: Information about reported content available	13
Outcome 1e: Information about recommender engines	14
Objective 2: Disrupt advertising and monetisation incentives for disinformation.	16
Objective 4: Empower consumers to make better informed choices of digital content.	19
Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.....	22
Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.....	23
Objective 7: Signatories publicise the measures they take to combat Disinformation.....	25
<i>Concluding remarks</i>	27
<i>Appendix</i>	28



Summary

Introduction

TikTok is [committed](#) to nurturing creativity in a safe, supportive and authentic environment. In a global community, it is natural for people to have different opinions, but we seek to operate on a shared set of facts and reality.

The [Integrity and Authenticity \(I&A\)](#) policies within our [Community Guidelines](#) prohibit harmful misinformation, impersonation, and coordinated or synthetic, manipulated misleading content. Violative videos are removed from the platform, and these efforts are detailed in TikTok's quarterly [Community Guidelines Enforcement Reports](#).

Key initiatives undertaken by TikTok to support long-term risk mitigation include:

- investment in machine learning models to ensure extensive coverage of nuanced misinformation threats.
- detection and removal of inauthentic visual and audio trends to help combat manipulated, edited and deepfake content.
- maintaining a database of fact-checked claims enabling human moderators to accurately identify misinformation content.
- launching an anti-misinformation program in partnership with accredited fact-checkers to assess inauthentic narratives on third party platforms and restrict them from TikTok.
- developing a first-of-its-kind [AI-generated content label](#) launched in September 2023 to help people identify any realistic AIGC, with stringent [rules](#) and new [technologies](#) to proactively address AIGC related misinformation.

We work with International Fact-Checking Network-accredited fact-checking partners to enforce our rules against harmful misinformation. This work is particularly important during elections and other civic processes, as it enables us to verify claims and take action in line with our Community Guidelines. As a precautionary measure, while content is subject to review by our fact-checking partners, it remains on the platform but is not eligible for recommendation on the For You Feed. This helps to ensure that content produced in the context of an election period can be thoroughly assessed by independent fact-checking partners before misinformation enforcement decisions are made. While this process may sometimes result in delayed content moderation decisions, depending on the context and the complexity of the topic, it is designed to reduce the risk of mismoderating legitimate political discourse. Further information on our work to manage the risks associated with harmful misinformation while supporting freedom of expression is set out in the Appendix.



Notable Highlights

In 2023, TikTok implemented a number of global and Australia-specific initiatives to combat misinformation and ensure a safe user experience. These included:

- Collaborating with the NSW Electoral Commission and the Australian Electoral Commission to support the New South Wales State Election and the Indigenous Voice to Parliament Referendum respectively.
- Providing Public Service Announcements (**PSAs**) for the NSW State Election through hashtags and search terms, reminding users of our Community Guidelines and directing users to the NSW Electoral Commission's website.
- Launching the 2023 Australian Referendum Hub, directing users to a dedicated page with authoritative information on the Referendum.
- Implementing new front-end product safety features for the Voice Referendum, such as a Search Guide and Notice Tags for both short-form videos and TikTok LIVE.
- Maintaining fact-checking partnerships with the Australian Associated Press (**AAP**) to prevent the spread of misleading information.
- Deploying additional Arabic- and Hebrew-speaking moderators in response to the Israel-Hamas war, and launching search interventions which are triggered when users search for terms related to this topic (e.g., "Israel", "Palestine").

Throughout the 2023 reporting period, we experienced fluctuations in our <24 hour video removal rate in both Q1 and Q3, as well as our removal before content receives any views rate in Q3 (Fig. 1A). This can in part be attributed to a concomitant increase in third-party fact-checking escalations, particularly regarding conspiracy theories including those related to the Indigenous Voice to Parliament Referendum. Fact-checking played a critical role in mitigating the spread of misinformation on-platform throughout the Referendum campaign period. In the 6-week period leading up to polling day on 14 October 2023, we escalated approximately 1,700 videos to fact-checkers, and enforced approximately 380 of them (i.e. removing the video from the For You feed or from the platform completely). Many of these videos were found to be propagating conspiracy theories, in violation of our Community Guidelines. Given the nuances of the Referendum debate, some of this escalated content required more extensive review to verify factual accuracy, which impacted our overall video removal metrics. These video removal rates subsequently stabilised in Q4, as represented in Fig. 1B below.

Fig. 1A below shows the volume of content removed for violating our harmful misinformation policies in Australia by quarter in 2023, as well as our performance metrics on removal efficiency.

Quarter (2023)	Total videos removed	Proactive removal ¹	Removal before content receives any views	Removal within 24 hours of content posted to platform
January - March	6,737	97.80%	76.90%	12.60%
April - June	10,721	96.30%	77.70%	71.50%
July - September	6,134	92.45%	43.85%	16.22%
October - December	4,919	94.20%	47.25%	41.74%

Fig. 1A: Summary of Removal of Harmful Misinformation Violations in 2023 (Australia)

Fig. 1B below shows that our proactive removal rates for Harmful Misinformation in Australia have consistently remained above 90% throughout 2023, as compared to proactive removal rates as low as 66.10% in Q1 2022.

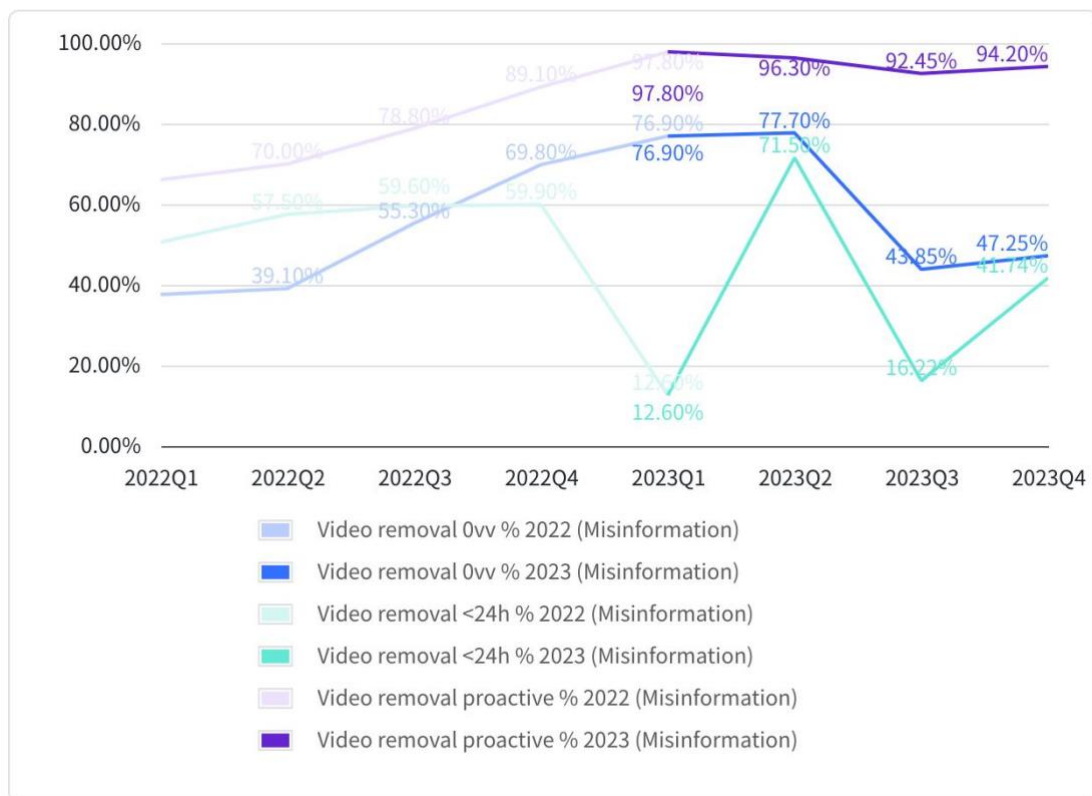


Fig. 1B: Removal of Integrity and Authenticity Violations Breakdown in 2023 vs 2022 (Australia)

¹ "Proactive removal" in this context refers to policy enforcement before it is reported by users.



TikTok's Commitments under the Code

TikTok opts in to all Objectives and Outcomes under the Australian Code of Practice on Disinformation and Misinformation with respect to the TikTok platform.

<p>Objective 1: Provide safeguards against Harms that may arise from Disinformation and Misinformation</p>	
<p><u>Outcome 1a:</u> Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures</p> <p><u>Outcome 1b:</u> Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code</p> <p><u>Outcome 1c:</u> Users can report content or behaviours to Signatories that violate their policies under section 5.10 through publicly available and accessible reporting tools.</p> <p><u>Outcome 1d:</u> Users will be able to access general information about Signatories' actions in response to reports made under 5.11.</p> <p><u>Outcome 1e:</u> Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.</p>	<p>Opt in (to all)</p>
<p>Objective 2: Disrupt advertising and monetisation incentives for Disinformation and Misinformation.</p>	
<p><u>Outcome 2:</u> Advertising and/or monetisation incentives for Disinformation and Misinformation are reduced.</p>	<p>Opt in</p>
<p>Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms</p>	
<p><u>Outcome 3:</u> The risk that Inauthentic User Behaviours undermine the integrity and security of services and products is reduced.</p>	<p>Opt in</p>
<p>Objective 4: Empower consumers to make better informed choices of digital content.</p>	
<p><u>Outcome 4:</u> Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation.</p>	<p>Opt in</p>

Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.	
<u>Outcome 5:</u> Users are better informed about the source of Political Advertising.	Opt in
Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.	
<u>Outcome 6:</u> Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation.	Opt in
Objective 7: Signatories publicise the measures they take to combat Disinformation and Misinformation.	
<u>Outcome 7:</u> The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.	Opt in

The following sections of the report will outline the specific measures, policies and projects undertaken to promote authenticity and counter misinformation on TikTok.





Reporting against 2023 commitments

Objective 1: Provide safeguards against Harms that may arise from Disinformation and Misinformation

Outcome 1a: Reducing harm by adopting scalable measures

In 2023 we implemented new scalable measures (outlined below) and strengthened our policy framework to help safeguard users from potential harms associated with misinformation. We regularly review these measures and ensure that people understand our guidelines, including what kind of content is not allowed on the platform, and when and why we may take action to mitigate any potential risks.

Our policies on misinformation are a subset of our Integrity and Authenticity policies. The 2023 reporting period saw an increase in Integrity and Authenticity policy violations as a proportion of total Community Guidelines violations, partly as a result of numerous significant civic processes, including the Indigenous Voice to Parliament Referendum, during which platforms typically experience higher volumes of misinformation.

Fig. 2A below shows the volume of content violating our Integrity and Authenticity policies in Australia by quarter in 2023, outlining the proportion of Integrity and Authenticity violations against all removed content, as well as our performance metrics on removal efficiency.

Quarter (2023)	Total videos removed	Proactive removal [1]	Removal before content receives any views	Removal within 24 hours of content posted to platform
January - March	8,764	95.90%	72.10%	29.50%
April - June	14,985	96.30%	80.90%	78.40%
July - September	10,672	93.30%	59.50%	48.40%
October - December	12,738	94.20%	70.90%	74.00%

Fig. 2A: Removal of Integrity & Authenticity Violations in 2023 (Australia)



Fig. 2B below outlines an increase in the proportion of Integrity and Authenticity Community Guidelines violations we detected on our platform in Australia in 2023, with Integrity and Authenticity violations accounting for a range of 0.90% - 1.70% of all Community Guidelines violations compared to a range of 0.60% - 0.90% in 2022. Despite the increase in proportion of Integrity and Authenticity violations, our proactive removal rate remains consistently high.

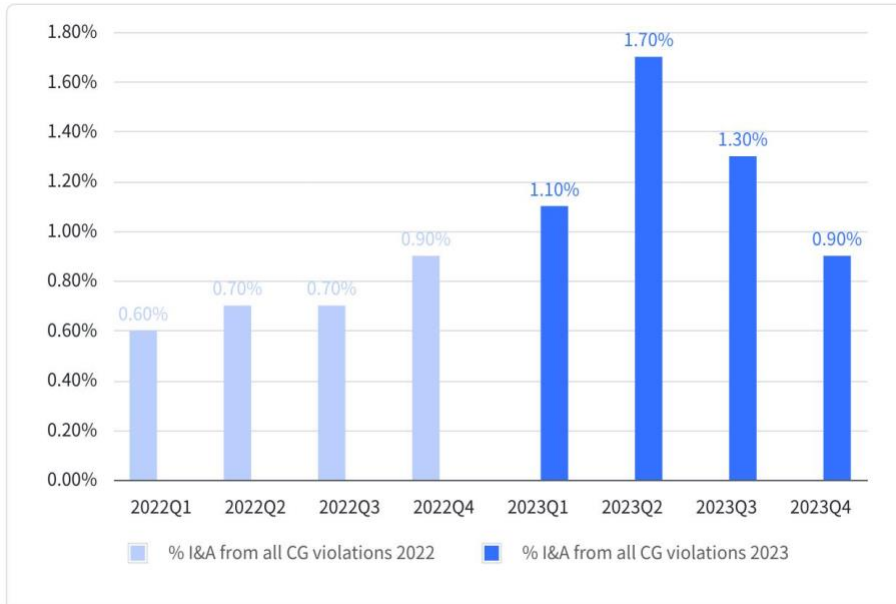


Fig. 2B: Proportion of Integrity and Authenticity violations in 2023 vs 2022 (Australia)

Fig. 2C below shows that our proactive removal rates for Integrity and Authenticity violations in Australia have consistently remained above 90% throughout 2023, as compared to proactive removal rates beginning at 83.60% in Q1 2022.

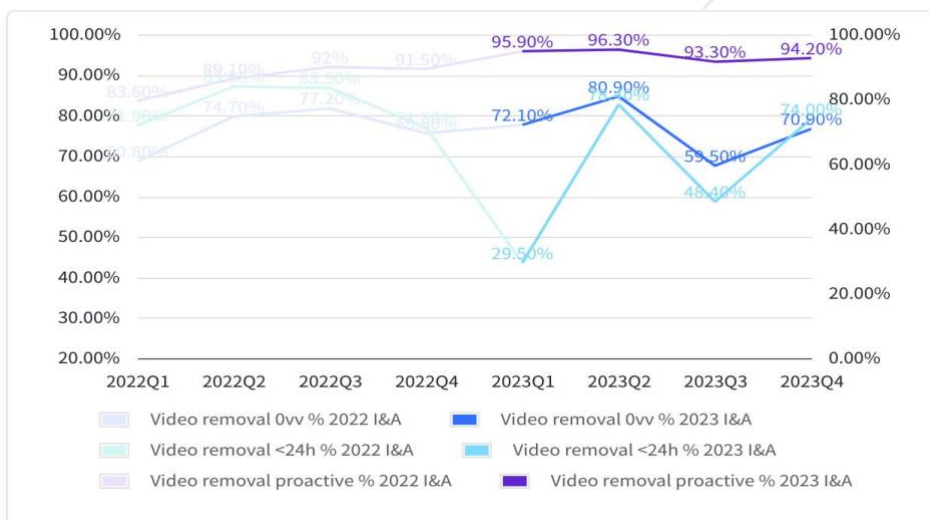


Fig. 2C: Removal of Integrity and Authenticity Violations Breakdown in 2023 vs 2022 (Australia)

New labels for disclosing AI-generated content

In September 2023, we introduced [new labels for users to disclose AI-generated content](#) (see Fig. 5) to make clear to viewers when content is significantly altered or modified by AI technology. The labels help creators showcase the innovations behind their content, and can be applied to any content that has been completely generated or significantly edited by AI.

This measure also makes it easier for users to comply with our Community Guidelines' [synthetic media policy](#), which we introduced in early 2023. The policy requires people to label AI-generated content that contains realistic images, audio or video, in order to help viewers contextualise the video and prevent the spread of potentially false, misleading, or deceptive content.

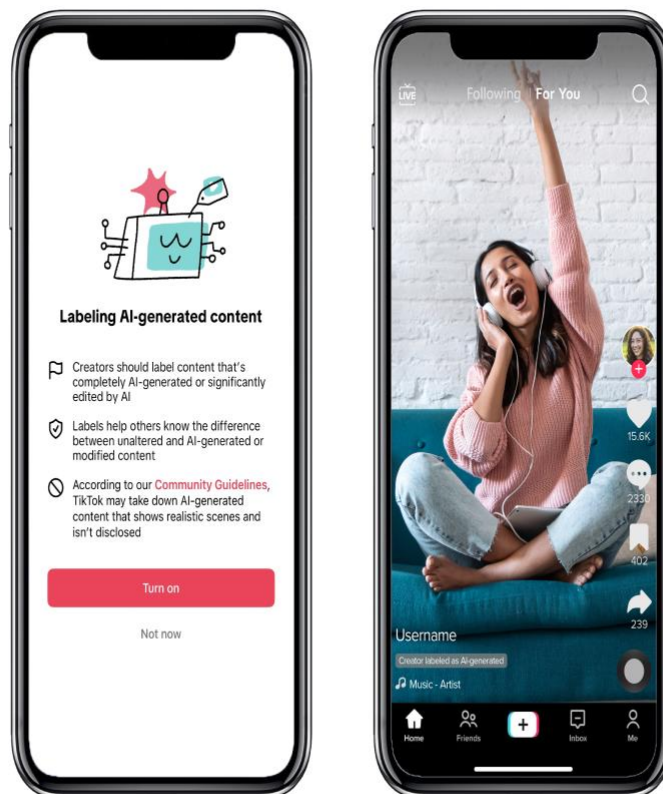


Fig. 3: AI Generated Content Labels

Strengthening enforcement through fact-checking partnerships

Globally, TikTok employs more than 40,000 Trust & Safety professionals who are responsible for ensuring the safety of the TikTok platform. This includes the development, implementation and enforcement of our harmful misinformation policies. During the 2023 reporting period, we collaborated



with 17 fact-checking organisations accredited by the International Fact-Checking Network (IFCN), covering over 50 languages, to ensure the accurate application of our Community Guidelines against misinformation (note: as of the date of reporting, TikTok currently collaborates with 18 such organisations). These partnerships empower our moderation teams to make accurate assessments of potentially misleading claims.

Our 24/7 moderation teams, along with our third-party fact-checking partners, work to review and verify flagged content and accounts. Throughout 2023 we continued to partner with the Australian Associated Press to help us independently review and assess the accuracy of content on our platform in Australia. Where such content is posted and is assessed to be false or deceiving, we remove such content in line with our [Community Guidelines](#), and where fact-checks are inconclusive, we may label and restrict the content from appearing in the For You feed, as detailed in the "[For You feed Eligibility Standards](#)" section of our Community Guidelines.

Outcome 1b: Inform users about what content is targeted

TikTok's [Community Guidelines](#) are available to users within the app and on our website, and includes detailed descriptions of what constitutes misinformation, what forms of harmful misinformation are not allowed on our platform, and the eligibility criteria for content to appear in users' feeds.

To more clearly inform users about what content constitutes mis/disinformation, in March 2023 we updated our Community Guidelines governing harmful misinformation and disinformation. The March 2023 update significantly expanded the descriptions of content we control within pursuant to our Integrity and Authenticity policies. The revised CGs expand upon our controls on 'Misinformation' more broadly, as well as material related to Civic and Election Integrity, Synthetic and Manipulated Media, and Fake Engagement. The updated CGs also noted that our policies target misleading content as well as that which is inaccurate and false content, and provided information on the type of content that we prevent from being promoted on the For You feed, but do not remove from the platform. This includes:

- General conspiracy theories that are unfounded and claim that certain events or situations are carried out by covert or powerful groups, such as “the government” or a “secret society”.
- Unverified information related to an emergency or unfolding event where the details are still emerging.
- Potential high-harm misinformation while it is undergoing a fact-checking review.
- We also clarified that we allow:



- Statements of personal opinion (as long as it does not include harmful misinformation)
- Discussions about climate change, such as the benefits or disadvantages of particular policies or technologies, or personal views related to specific weather events (as long as it does not undermine scientific consensus)

Specifically in relation to Election Misinformation, we included additional guidance that misinformation related to the following will constitute a violation of our Community Guidelines:

- How, when, and where to vote or register to vote.
- Eligibility requirements of voters to participate in an election, and the qualifications for candidates to run for office.
- Laws, processes, and procedures that govern the organisation and implementation of elections and other civic processes, such as referendums, ballot propositions, and censuses.
- The outcome of an election.

These policies also prevent from promotion on the For Your feed any content containing unverified claims about the outcome of an election that is still unfolding and may be false or misleading. For more information about our Integrity & Authenticity policies, please refer to the Appendix.

Our Community Guidelines are informed through extensive consultations with relevant stakeholders, including NGOs, regulators, academics, subject matter experts, as well as our community. We also ensure that these are regularly reviewed, and where appropriate, updated, and that our community is notified of any major changes.

Outcome 1c: Users can easily report offending content

TikTok is designed so that users can easily report content they consider to be potentially violative of our Community Guidelines. This includes a dedicated category to report misinformation and a selection of sub-categories to choose from. In 2023 we made refinements to these subcategories to include election misinformation, harmful misinformation, as well as deepfake, synthetic media, and manipulated media.

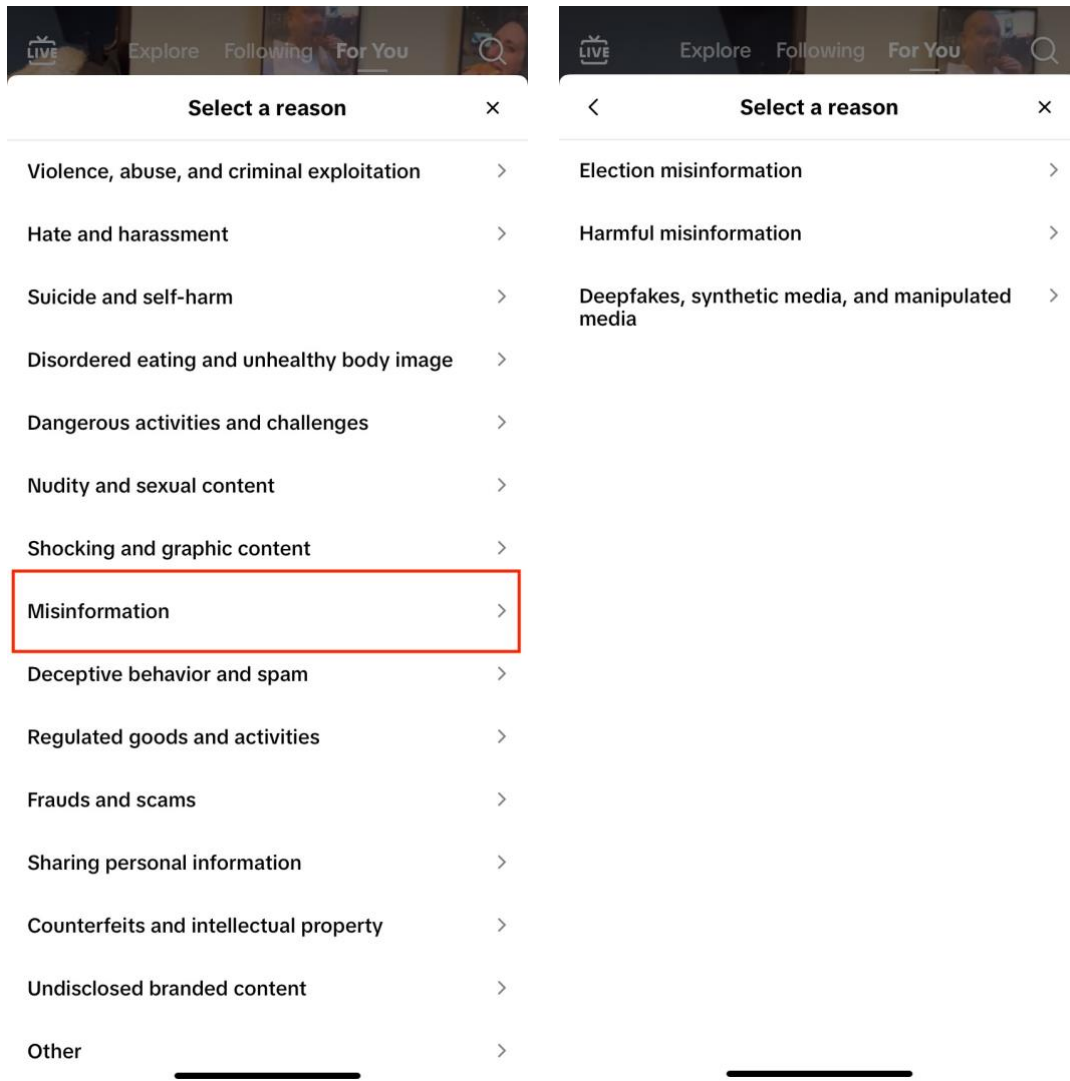


Fig. 4A: User in-app reporting interface

Before a user reports a type of misinformation to us, we make clear what we don't allow under each specific subcategory. This aims to enhance user awareness of our Community Guidelines, reduce ambiguity about permitted content, and strengthen platform safety.

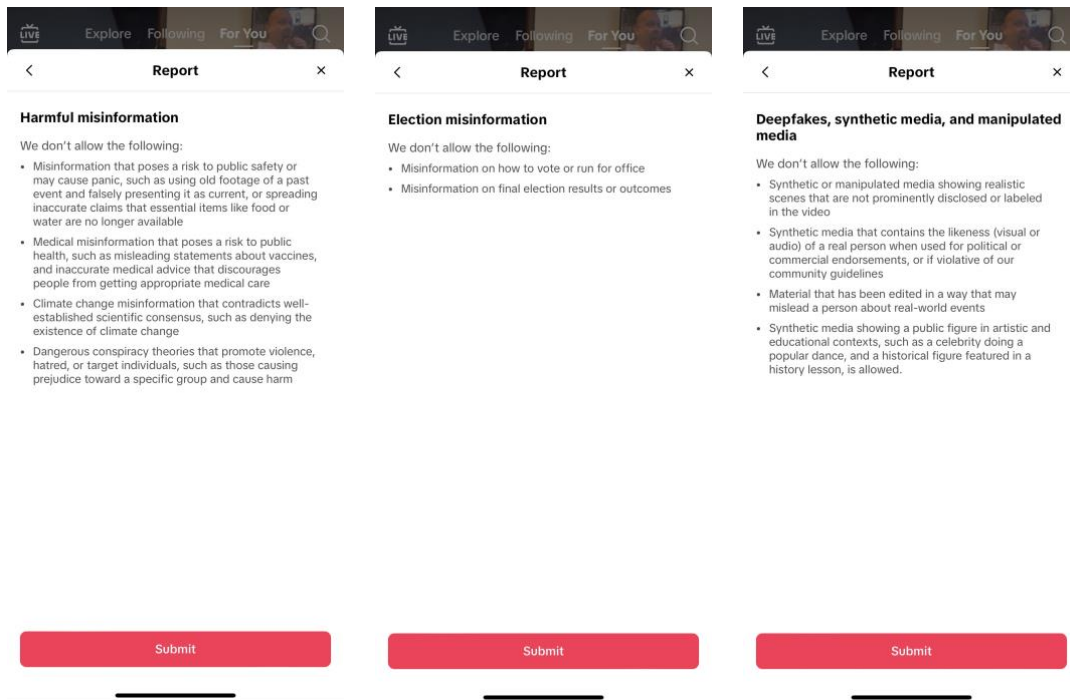


Fig. 4B: User in-app reporting interface

Users can keep track of their reports (including their status) and view their report history under Settings and Privacy > Support > Safety Center. When a user submits a report, we also provide them with the option to hide any further content being shown from the account in the feed.

Reporting misinformation is not limited to short-form videos. We enable users to also report misinformation across other features of the platform, including [comments](#) on videos, [direct messages](#) they receive from other users, [accounts](#), [sounds](#), [hashtags](#) and [auto-suggestions](#) generated when they search for something on TikTok. Users can also report [LIVE videos](#) and [comments](#) on livestreams if they encounter content that violates our [Community Guidelines](#).

We also have reporting channels for non-users to report potentially harmful material. Our reporting form hosted on our website enables direct reports for our immediate review and action. Instructions for our publicly accessible reporting tools are available on our [website](#).

Outcome 1d: Information about reported content available

Our Transparency Centre serves as a central hub to understand how TikTok moderates content, develops products, and protects user data. It provides users and the broader public with access to data and periodic reports, including:

- [Community Guidelines Enforcement Reports](#): Quarterly insights into our efforts to enforce guidelines and terms of service.



- [Information Requests Reports](#): Biannual data on user information requests from governments and law enforcement, along with our responses.
- [Government Removal Requests Reports](#): Biannual data on requests from government agencies to restrict content and our actions in response.
- [Intellectual Property Removal Requests Reports](#): Biannual data on requests to remove copyrighted and trademarked content, along with our responses.

These reports are published in multiple languages, are available for download in machine-readable formats from our Transparency Centre, and can be visualised in interactive charts and graphs.

Our [latest Community Guidelines Enforcement Report](#) for the period October 2023 - December 2023, published in March 2024, summarises our capabilities to proactively detect and remove violative material from our platform in Australia.

Quarter (2023)	Total videos removed	Proactive removal rate	Removal before content receives any views	Removal rate within 24 hours
January - March	792,784	95.60%	65.00%	81.10%
April - June	832,120	96.50%	77.70%	88.70%
July - September	741,846	95.40%	70.10%	85.50%
October - December	1,222,046	95.70%	74.70%	88.30%

Fig. 5: Video Removals in 2023 (Australia)

Outcome 1e: Information about recommender engines

The content people see on TikTok is generated by our community and recommendations are based on the content people have previously engaged with. Using signals such as view counts, likes, and shares, the recommendation algorithm creates a prediction score to rank videos to potentially recommend.

Our [support page](#) provides detailed information to users about how content is recommended across TikTok and how users can influence what they see on the platform. We have also provided additional information in our [Transparency Centre](#).

Aside from the signals a user provides by how they interact with content on TikTok, there are additional tools we have built to help our community better control what kind of content is recommended to them. These include:



- **Not interested:** A user can long-press on the video in your For You feed and select '[Not interested](#)' from the pop-up menu. This will let us know they are not interested in this type of content and we will limit how much of that content we recommend.
- **Video keyword filters:** A user can [add keywords](#) – both words or hashtags – they'd like to filter from their For You feed.
- **For You feed refresh:** To help discover new content, a user can [refresh the For You feed](#), which provides an entirely new side of TikTok for them to explore.



Objective 2: Disrupt advertising and monetisation incentives for disinformation.

We place considerable emphasis on proactive content moderation and the vast majority of the violative content we remove is taken down before it is reported to us or receives any views. We are also committed to continuing to keep pace with evolving issues that affect our users.

Our work since the last report continues to reflect our strong commitment to combatting disinformation on our platform and to providing transparency to our wider community about the measures we take. We are in the process of developing and launching more granular misinformation policies and policies to govern AI-generated content in the coming months.

Transparency and Scrutiny of Advertising

Ads must comply with and are reviewed against our [ad policies](#) before being allowed on our platform. These policies specifically prohibit misleading, inauthentic and deceptive behaviours.

We continue to engage with external stakeholders in order to increase the effectiveness of our scrutiny of ad placements. As a Global Alliance for Responsible Media (GARM) member, we also remain committed to upholding the GARM Framework and, as part of that, removing harmful misinformation from monetisation.

Like all users of our platform, participants in content monetisation programs must adhere to our [Community Guidelines](#), including our Integrity and Authenticity policies. Those policies make clear that we do not allow activities that may undermine the integrity of our platform or the authenticity of our users. They also make clear that we remove content or accounts, including those of creators, which contain misleading information that causes significant harm or deceptive behaviours. In certain scenarios, we may remove a creator's access to a creator monetisation feature.

Our policies and approach

Our [Integrity and Authenticity](#) policies within our [Community Guidelines](#) are the first line of defence in combating harmful misinformation and deceptive behaviours on our platform.

Paid ads are also subject to our [ad policies](#) and are reviewed against these policies before being allowed on our platform. Our ad policies specifically prohibit inaccurate, misleading, or false content that may cause significant harm to individuals or society, regardless of intent. They also prohibit other misleading, inauthentic and deceptive behaviours. Ads deemed in violation of these policies will not



be permitted on our platform, and accounts deemed in severe or repeated violation may be suspended or banned.

We also have other, existing ad policies that focus on certain topics where the risk of disinformation may be higher. By way of example, our [Covid-19 ad policy](#) prohibits ads that seek to take advantage of Covid-19 to push sales, for example by manipulating consumers' fear or anxiety, or spreading harmful misinformation to push sales. As well as ensuring ads relating to Covid-19 do not spread harmful misinformation, we also promote authoritative sources of information.

We are continually reflecting on whether there are further focused areas for which we should develop new policies. Our [ad policies](#) require advertisers to meet a number of requirements regarding the landing page. For example, the landing page must be functioning and must contain complete and accurate information, including about the advertiser. Ads may not be approved if the product or service advertised on the landing page does not match that included in the ad.

We make various brand safety tools available to advertisers to assist in helping to ensure that their ads are not placed adjacent to content they do not consider to fit with their brand values. While any content that is violative of our Community Guidelines, including our Integrity & Authenticity policies, is removed, the brand safety tools are designed to help advertisers to further protect their brand. As a GARM member, we believe in its mission and have adopted GARM's Brand Safety Floor and Suitability Framework (the **GARM Framework**).



Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms.

TikTok remains committed to preventing, detecting and deterring inauthentic user behaviours on our platform. These efforts include removing inauthentic accounts, tackling fake account engagement and disrupting [Covert Influence Operations \(CIO\)](#). In 2023, we disrupted a total of 46 networks globally.

We have enhanced our ability to detect CIO through a dual-pronged strategy that focuses on enhancing detection efforts by integrating insights gained from extensive global investigations, as well as developing strategic partnerships with third-party intelligence providers to complement existing in-house capabilities. We also consult with members of our Safety Advisory Council to gain insight into, and obtain advice on, our efforts to detect covert influence operations as we continually work to improve our efforts in this regard.

Where our teams have a high degree of confidence that an account is engaged in inauthentic coordination of content creation or amplification, uses deceptive practices to deceive/manipulate platform algorithms or coordinated mass reporting of non-violative opposing content/accounts and is engaged in or is connected to networks we took down in the past as part of a CIO, it is removed from our Platform in accordance with our CIO policy.

When we investigate and remove these operations, we focus on behaviour and assessing linkages between accounts and techniques to determine if actors are engaging in a coordinated effort to mislead TikTok's systems or our community. We know that CIOs will continue to evolve in response to our detection and networks may attempt to re-establish a presence on our platform. We continue to iteratively research and evaluate complex deceptive behaviours on our platform and develop appropriate product and policy solutions as appropriate in the long term. We publish the details of all of the CIO networks we identify and remove within our transparency reports, [here](#).

[Our Integrity & Authenticity policies](#), which address fake engagement, do not allow the trade of services that attempt to artificially increase engagement or deceive TikTok's recommendation system. We do not allow our users to facilitate the trade of services that artificially increase engagement, such as selling followers or likes, or to provide instructions on how to artificially increase engagement on TikTok.

If we become aware of accounts or content with inauthentically inflated metrics, we will remove the associated fake followers or likes. Content that tricks or manipulates others as a way to increase engagement metrics, such as "like-for-like" promises and false incentives for engaging with content

is ineligible for our For You feed. To know more about our approach on disrupting CIO, please refer to the Appendix.

Objective 4: Empower consumers to make better informed choices of digital content.

We take a multi-pronged approach to enabling users to make better informed choices of content on TikTok including implementing in-app features to provide timely, accurate and authoritative information to users regarding major civic events such as elections and referenda, and content relating to public health issues.

Combatting the spread of misinformation for the Indigenous Voice to Parliament Referendum

In light of the national significance and importance of the Voice to Parliament Referendum in October 2023, TikTok dedicated additional Trust and Safety resources to enact a broad spectrum of mitigative measures to detect and prevent the spread of harmful referendum-related misinformation during the Referendum campaign.

Our teams relied on both internal and external intelligence sources to address general misinformation trends, including ones logged in the [Australian Electoral Commission's Disinformation Register](#). This ensured our teams were able to effectively identify and act upon known misinformation trends.

In addition, we created a Referendum Hub in collaboration with SBS's National Indigenous Television channel (**NITV**) and implemented front-end product features such as a Search Guide² and a Notice Tag.³ These features ensured our users who were seeking or viewing content associated with the Referendum were able to access authoritative and factual sources of information from community partners and the Australian Electoral Commission.

² An information panel that appears when users search referendum-related terms which briefly explains what the Referendum is about. When the information panel is clicked, it will lead users to the Referendum Hub.

³ A small text-based banner that appears on the bottom of videos associated with the Referendum. When the text based banner is clicked, it will lead users to the Referendum Hub.

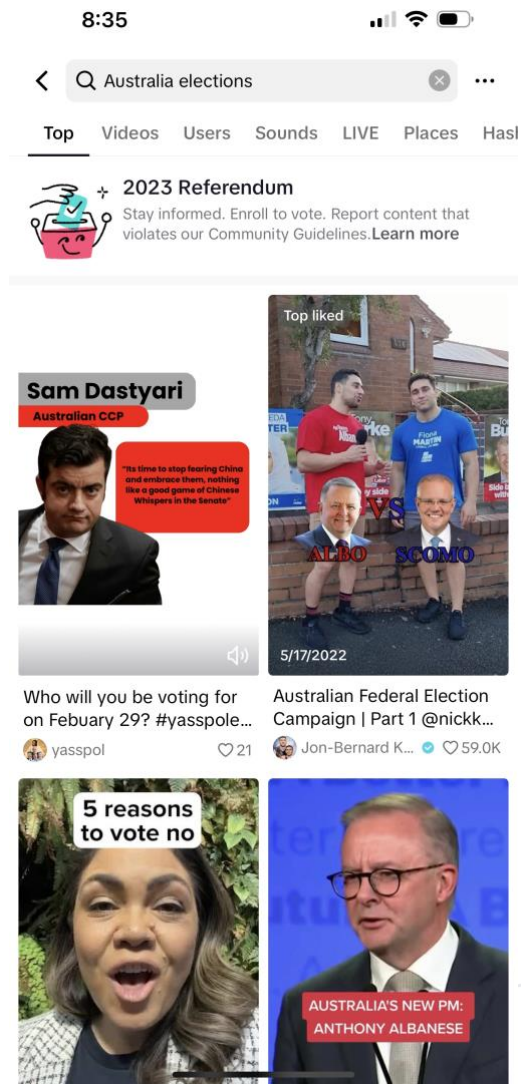


Fig. 6: Referendum Hub (Left), Search Guide (Right)

These measures saw strong engagement from our community. During this time:

- our Search Guide received approximately 350,000 impressions.
- our Referendum Hub was viewed approximately 76,000 times.
- Notice Tags were applied to approximately 191,000 TikTok video and photo posts, with approximately 49.5 million impressions.
- Notice Tags applied to TikTok LIVE-related content received approximately 400,000 impressions.



Fact-checking played a critical role in mitigating the spread of misinformation on-platform throughout the referendum campaign period. In the 6-week period leading up to polling day on 14 October 2023, we escalated approximately 1,700 videos to fact checkers, and enforced approximately 380 of them (i.e. removing the video from the For You feed or from the platform completely). Many of these videos were found to be propagating conspiracy theories, in violation of our Community Guidelines.



Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.

Transparency and Scrutiny of Advertising

Like all users of our platform, participants in content monetisation programs must adhere to our [Community Guidelines](#), including our Integrity and Authenticity policies. These policies include our prohibition on paid political advertising.

Prohibiting Paid Political Ads

TikTok does not allow anyone to place [political ads](#), nor do we allow politicians and political party accounts to place ads. We also prevent [Government, Politician, and Political Party Accounts \(GPPAs\)](#) from accessing our monetisation features and campaign fundraising.

Sharing political beliefs and engaging in political conversation is allowed as organic content, but our policies prohibit users from paying to advertise or promote this content. We allow some cause-based advertising and public service advertising from government agencies, non-profits and other entities, provided they are not politically partisan and make exceptions for governments in certain circumstances, e.g., to promote public health.

Transparency Risk Controls

Where accounts are designated as GPPAs, those accounts are banned from placing ads on TikTok (with the exception of certain government entities in certain circumstances, as outlined above) and from monetisation features. We publish the details of our GPPA policy on our [website](#), where we set out who we consider to be a GPPA and the restrictions on those types of accounts.

We apply an internal label to accounts belonging to a [government, politician, or political party](#). Once an account has been labelled in this manner, a number of policies will be applied that help prevent misuse of certain features, e.g., access to advertising features and solicitation for campaign fundraising are not allowed.



Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.

We recognise the important role of researchers and subject matter experts in helping to identify misinformation trends and practices.

Advancing our commitment to combat climate misinformation

We remain committed to increasing climate literacy among our global community. As outlined above, climate change misinformation is directly referenced within our Community Guidelines. We permit discussions about climate change, such as the benefits or disadvantages of particular policies or technologies, or personal views related to specific weather events (as long as it does not undermine scientific consensus), but do not allow climate change misinformation that undermines well-established scientific consensus, such as denying the existence of climate change or the factors that contribute to it.

As part of our ongoing commitment, in November 2023, to coincide with the [COP28](#) UN Climate Change Conference, we launched our [#ClimateAction campaign](#) with new initiatives and programming. This included the introduction of a \$1 million initiative to tackle climate misinformation in support of [Verified for Climate](#), a joint program of the United Nations and Purpose. The initiative helps bring together a network of climate messengers with credibility and experience (known as 'Verified Champions)', including scientists and trusted experts from Brazil, the United Arab Emirates, and Spain who support select TikTok creators in developing educational content to tackle climate misinformation and disinformation.

Ongoing support with Australian Associated Press Fact Check

During the 2023 referendum period, Trust & Safety staff also participated in a closed event hosted by our fact-checking partners, Australian Associated Press, and stakeholders, where representatives of different platforms discussed their respective approaches to content moderation for issues like misinformation and listened to commentary and questions concerning the fact-checking process.

Community Partner Channel

Our Global [Community Partner Channel](#) provides selected organisations an additional route for reporting content that they believe breaks our Community Guidelines so that it can be reviewed by



our teams. To date, more than 400 organisations who specialise in a range of safety issues use our Community Partner Channel. In Australia, 25 partners have been introduced to the program, including organisations focusing on combating antisemitism, Islamophobia, hate speech and racism.

Objective 7: Signatories publicise the measures they take to combat Disinformation.

TikTok remains committed to transparency and to ensuring the clarity of our practices for users, law enforcement agencies, governments and the general public. We continue active engagements with government and regulatory bodies, providing visits to our Transparency and Accountability Centres (**TACs**) to see up close how we moderate and recommend content, secure our platform, and protect people's privacy. Our newly opened TAC in Singapore complements our existing TACs in Dublin and Los Angeles, and has hosted hundreds of guests both physically and virtually.

We will continue our efforts to publish quarterly transparency reports and provide newsroom updates. In 2023, we broadened the scope of information included in our Community Guidelines Enforcement Reports, which include reporting the number of removed suspected under-age accounts and information on the identification and removal of covert influence operations globally.

Helping our community access authoritative information on the Israel-Hamas war

In the aftermath of the start of the Israel-Hamas war, we immediately mobilised significant resources and personnel to help maintain the safety of our community and integrity of our platform. Our focus has been on supporting free expression, upholding our commitment to human rights, and maintaining the safety of our community and integrity of our platform during the war. We have provided regular updates on our efforts as part of our commitment to transparency. In the six months since October 7, 2023, we have removed more than 3.1 million videos and suspended more than 140,000 livestreams in Israel and Palestine for violating our Community Guidelines, including content promoting hate speech, violent extremism and misinformation.⁴ Approximately 155,000 videos propagating dangerous misinformation have been removed globally.

We are continually working hard to ensure that TikTok is a source of reliable and safe information and recognise the heightened risk and impact of misleading information during a time of crisis. As part of our crisis management process, we launched a command centre that brings together key members of our 40,000-strong global team of safety professionals, representing a range of expertise and regional perspectives, so that we remain agile in how we take action to respond to this fast-evolving crisis.

⁴ Data covers October 7, 2023 to March 31, 2024.



Since the onset of the war, there has been a rise in misinformation and conspiracy theories relating to the war. We have also seen spikes in deceptive account behaviours and continue to take swift action against fake engagement and accounts, for example, by removing 35 million fake accounts in the month after the start of the war - a 67% increase in the previous month. From October 7 through to the end of 2023, we have removed more than 169 million fake accounts globally, as well as removing approximately 1.2 million bot comments on content tagged with hashtags related to the war.

To help raise awareness and to protect our users, we have also launched search interventions which are triggered when users search for non-violating terms related to the war (e.g., Israel, Palestine). These search interventions remind users to pause and check their sources and also direct them to wellbeing resources.

We remain committed to engagement with experts across the industry and civil society, such as our Safety Advisory Councils, and cooperation with law enforcement agencies globally in line with our Law Enforcement Guidelines, to further safeguard and secure our platform during these difficult times.



Concluding remarks

TikTok is committed to upholding its obligations under the Australian Code of Practice for Disinformation and Misinformation. This report has highlighted our continued efforts - both locally and globally - to combat misinformation through a range of new and strengthened measures and policy frameworks. We remain dedicated to safeguarding our users from the risks associated with harmful misinformation.

We recognise that misinformation is an evolving issue, and with new tools and advancements in technology, including with respect to generative AI, we are committed to proactively enhancing our methods of detecting and mitigating risks associated with such content, as well as assessing options for partnerships to determine the origin of content. We also continue to partner closely with experts to ensure we are able to stay ahead of emerging trends, and effectively mitigate risks associated with AI-related misinformation.

As TikTok's user base continues to grow, we strive to enable creativity in a safe environment, as well as to support genuine discussions on global affairs, politics and health. We deeply value the trust of our community and are committed to providing a transparent, authentic platform experience globally.



Appendix

Approach to Disinformation and Misinformation

Our misinformation policies apply to content regardless of the poster's intent, as the content's harm is the same either way. Hence, they cover both "disinformation" (which is intentionally shared to mislead) and harmful misinformation that may not have been shared with the goal of deceiving people.

Like others in our industry, we do not prohibit people from sharing personal experiences, simply inaccurate myths, or misinformation that could cause reputational or commercial harm, in order to balance creative expression with preventing harm.

Policy on Misinformation

In a global community, it is natural for people to have different opinions, but we seek to operate on a shared set of facts and reality. **We do not allow inaccurate, misleading, or false content that may cause significant harm to individuals or society, regardless of intent.** Significant harm includes physical, psychological, or societal harm, and property damage. It does not extend to commercial and reputational harm, nor does it cover simply inaccurate information and myths. We rely on [independent fact-checking partners](#) and our database of previously fact-checked claims to help assess the accuracy of content.

Content is ineligible for the FYF if it contains general conspiracy theories or unverified information related to emergencies. To be cautious, content that warrants fact-checking is also temporarily ineligible for the FYF while it is undergoing review.

To help you manage your TikTok experience, we add warning labels to content related to unfolding or emergency events which have been assessed by our fact-checkers but cannot be verified as accurate, and we prompt people to [reconsider sharing](#) such content.

Misinformation includes inaccurate, misleading, or false content.

Significant harm includes severe forms of:

- Physical injury and illness, including death
- Psychological trauma
- Large-scale property damage
- Societal harm, including undermining fundamental social processes or institutions, such as democratic elections, and processes that maintain public health and public safety



Conspiracy theories are beliefs about unexplained events or involve rejecting generally accepted explanations for events and suggesting they were carried out by covert or powerful groups.

NOT allowed

- Misinformation that poses a risk to public safety or may induce panic about a crisis event or emergency, including using historical footage of a previous attack as if it were current, or incorrectly claiming a basic necessity (such as food or water) is no longer available in a particular location
- Medical misinformation, such as misleading statements about vaccines, inaccurate medical advice that discourages people from getting appropriate medical care for a life-threatening disease, and other misinformation that poses a risk to public health
- Climate change misinformation that undermines well-established scientific consensus, such as denying the existence of climate change or the factors that contribute to it
- Dangerous conspiracy theories that are violent or hateful, such as making a violent call to action, having links to previous violence, denying well-documented violent events, and causing prejudice towards a group with a protected attribute
- Specific conspiracy theories that name and attack individual people
- Material that has been edited, spliced, or combined (such as video and audio) in a way that may mislead a person about real-world events

FYF ineligible

- General conspiracy theories that are unfounded and claim that certain events or situations are carried out by covert or powerful groups, such as “the government” or a “secret society”
- Unverified information related to an emergency or unfolding event where the details are still emerging
- Potential high-harm misinformation while it is undergoing a fact-checking review

Allowed

- Statements of personal opinion (as long as it does not include harmful misinformation)
- Discussions about climate change, such as the benefits or disadvantages of particular policies or technologies, or personal views related to specific weather events (as long as it does not undermine scientific consensus)



Since the start of the war on 7 October 2023, we have been working diligently to remove content that violates our policies. We have set out below some of the main threats both observed and considered in relation to the war and the actions we have taken to address these.

(I) Spread of misinformation

We believe that trust forms the foundation of our community, and we strive to keep TikTok a safe and authentic space where genuine interactions and content can thrive. TikTok takes a multi-faceted approach to tackling the spread of harmful misinformation, regardless of intent. This includes our: [Integrity & Authenticity policies](#) (I&A policies) in our [Community Guidelines](#) (CGs); as well as our external partnerships with fact-checkers, media literacy bodies, and researchers. We support our moderation teams with detailed misinformation policy guidance, enhanced training, and access to tools like our global database of previously fact-checked claims from our IFCN-accredited fact-checking partners, who help assess the accuracy of content.

Since 7 October 2023, there has been a rise in misinformation and conspiracy theories relating to the war. We have also seen spikes in deceptive account behaviours and continue to take swift action against fake engagement and accounts, for example, by removing 35 million fake accounts in the month after the start of the war - a 67% increase on the previous month.

(II) Covert Influence Operations (CIO)

TikTok's I&A policies do not allow deceptive behaviour that may cause harm to our community or society at large. This includes coordinated attempts to influence or sway public opinion while also misleading individuals, our community, or our systems about an account's identity, approximate location, relationships, popularity, or purpose. We have specifically-trained teams on high alert to investigate CIO and we provide quarterly updates on the CIO networks we detect and remove from our platform in our [Community Guidelines Enforcement Reports](#) (CGER).

We have assigned dedicated resourcing within our specialist teams to proactively monitor for CIO in connection with the war. While we have not identified any CIO specifically targeting the war during this reporting period, we reported on a CIO relevant to the region in our Q1 (Jan-March 2023) [CGER](#) where we identified and removed a network operated from Israel that targeted Israeli audiences. We are currently investigating a number of operations and will publish details of any CIO networks we identify and remove. While we currently report the removals of covert influence networks in the quarterly CGER, in the coming months, we will also introduce dedicated CIO reports to further increase transparency, accountability, and cross-industry sharing.

We know that CIOs will continue to evolve in response to our detection and networks may attempt to re-establish a presence on our platform, which is why we continually seek to strengthen our policies



and enforcement actions in order to protect our community against new types of harmful misinformation and inauthentic behaviours.



Australian Code of Practice on Disinformation and Misinformation

Twitch

Baseline Transparency Report - May 2024

Summary

Twitch is a live streaming service, where creators engage in a wide variety of different activities, including video games, music, cooking, and creating creative content.

At Twitch, we strive to create a space that supports and sustains streamers' ability to express themselves, and provides a welcoming and entertaining environment for viewers, free of illegal and harmful interactions. This starts with Twitch's Community Guidelines, which seek to balance user expression with community safety, and set the rules for the behaviour of everyone on Twitch. Our Community Guidelines are developed by a dedicated team of policy professionals in consultation with external safety, human rights, and policy experts, and we review and update them regularly to respond to the community's evolving needs.

We identify potential violations of our Community Guidelines using a combination of machine detection, proactive human review, and user reporting. Our global safety operations team works to quickly review content and accounts flagged by users and by our machine detection models. The speed at which we can respond to user reports is critical given the live nature of Twitch, and in H2 2023, we responded to 95% of reports in under 1 hour and 99.95% of reports in under 24 hours. We prioritise having a human in the review process to ensure that decisions are accurate and fair for our community members.

We take pride in how Twitch fosters community and brings people together, and we believe that individuals who use online services to spread false, harmful information do not have a place in our community. This is why we have a Harmful Misinformation Actor policy. Harmful misinformation actors account for a disproportionate amount of damaging, widely debunked misinformation online. These actors share three characteristics: their online presence—whether on or off Twitch—is dedicated to (1) persistently sharing (2) widely disproven and broadly shared (3) harmful misinformation topics, such as conspiracies that promote violence. We prohibit harmful misinformation actors who meet all three of these criteria since taken together they create the highest risk of harm, including inciting real world harm.

Even if someone is not a Harmful Misinformation Actor, Twitch prohibits and enforces against misinformation that targets specific communities under our Hateful Conduct & Harassment policies, and we take action on content that encourages others to engage in physically harmful behaviour under our Self-Destructive Behaviour policy.

In addition to misinformation, Twitch invests significant resources to ban bots, spammers, impersonators, and other types of bad actors to combat inauthenticity on our service. We have proactive detection working alongside our reporting system to programmatically remove bots, known bad actors, and those who are trying to evade a suspension or ban.

While misinformation is not currently prevalent on Twitch, we recognize the harm that this content can cause, particularly when it is related to an election. We are always evolving our approach to safety in accordance with expert guidance and trends in our community. We understand that the prevalence of harmful misinformation can change, and we will continue to engage with industry, academia, and civil society to adapt our approach as necessary to ensure its continuing effectiveness. We participate in a variety of industry knowledge-sharing initiatives—including the EU Code of Practice on Disinformation, the New Zealand Code of Practice for Online Safety and Harms (which also addresses disinformation), the EU Hate Speech Code, the EU Internet Forum, and the Global Internet Forum to Counter Terrorism (GIFCT)—to stay abreast of industry trends and risks.

We are proud to contribute to the goals and commitments of the Australian Voluntary Code of Practice on Disinformation and Misinformation (ACPDM).

Commitments under the Code

Twitch has committed to the following six Objectives and related Outcomes.

Objective 1 - Provide safeguards against harms that may arise from disinformation and misinformation	
1a	Signatories contribute to reducing the risk of harms that may arise from the propagation of disinformation and misinformation on digital platforms by adopting a range of scalable measures.
1b	Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.
1c	Users can report content or behaviours to Signatories that violate their policies under section 5.10 through publicly available and accessible reporting tools.
1d	Users will be able to access general information about Signatories' actions in response to reports made under 5.11.
1e	Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.
Objective 2 - Disrupt advertising and monetisation incentives for disinformation	
2	Advertising and/or monetisation incentives for disinformation and misinformation are

	reduced.
Objective 3 - Work to ensure the integrity and security of services and products delivered by digital platforms.	
3	The risk that Inauthentic User Behaviours undermine the integrity and security of services and products is reduced.
Objective 4 - Empower consumers to make better informed choices of digital content.	
4	Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify misinformation.
Objective 6 - Strengthen public understanding of disinformation and misinformation through support of strategic research	
6	Signatories support the efforts of independent researchers to improve public understanding of disinformation and misinformation.
Objective 7 - Signatories will publicise the measures they take to combat disinformation and misinformation.	
7	The public can access information about the measures Signatories have taken to combat disinformation and misinformation.

Twitch did not subscribe to Objective 5 (Improve public awareness of the source of political advertising carried on digital platforms) as we do not permit political ads.

Reporting against commitments

Outcome 1a: Reducing harm by adopting scalable measures

In order to reduce harm to our community and the public without undermining our streamers’ open dialogue with their audiences, we prohibit [Harmful Misinformation Actors](#) who persistently share misinformation on or off of Twitch. We suspend users whose online presence is dedicated to (1) persistently sharing (2) widely disproven and broadly shared (3) harmful misinformation topics.

This policy is focused on Twitch users who persistently share harmful misinformation, including AI-generated misinformation. This focus represents a refinement of the ACPDM’s definition of misinformation. Twitch’s policy will not be applied to users based upon individual statements or discussions that occur on the channel. We evaluate whether a user violates the policy by assessing both their on-service behaviour as well as their off-service behaviour.

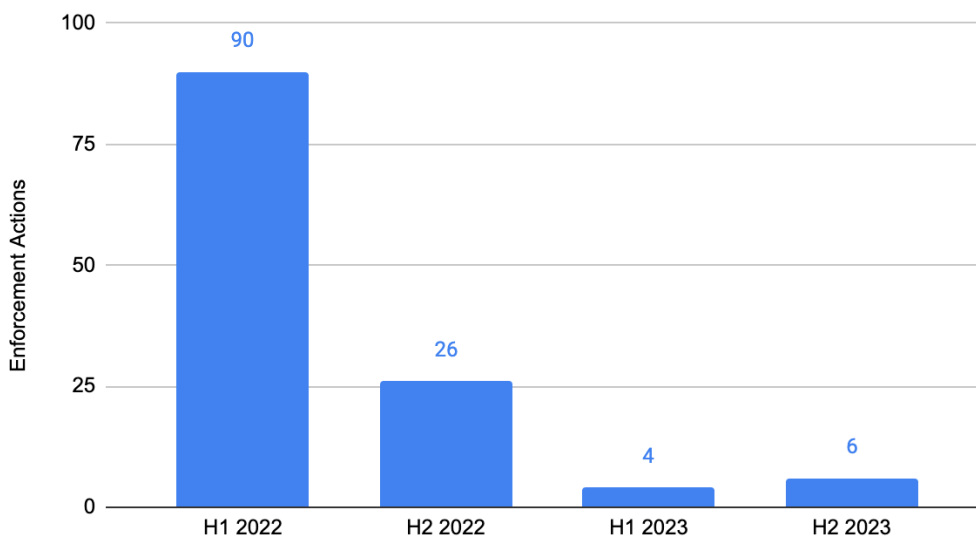
Under our Harmful Misinformation Actor Policy, we cover the following topic areas, and will continue to update this list as new trends emerge:

- Misinformation that targets protected groups, which is already prohibited under our Hateful Conduct & Harassment Policy

- Harmful health misinformation and wide-spread conspiracy theories related to dangerous treatments, COVID-19, and COVID-19 vaccine misinformation
 - Discussions of treatments that are known to be harmful without noting the dangers of such treatments
 - For COVID-19—and any other WHO-declared Public Health Emergency of International Concern (PHEIC)—misinformation that causes imminent physical harm or is part of a broad conspiracy
- Misinformation promoted by conspiracy networks tied to violence and/or promoting violence
- Civic misinformation that undermines the integrity of a civic or political process
 - Promotion of verifiably false claims related to the outcome of a fully vetted political process, including election rigging, ballot tampering, vote tallying, or election fraud
- In instances of public emergencies (e.g., wildfires, earthquakes, active shootings), we may also act on misinformation that may impact public safety)

In H2 2023, we indefinitely suspended 6 accounts globally for violating our Harmful Misinformation Actor Policy. Historical enforcement information is included in the chart below (we introduced our Harmful Misinformation Actor policy in H1 2022).

Misinformation Global Enforcements



Our enforcement numbers are relatively low due to several factors. (i) The mechanics of Twitch are not conducive to spreading misinformation or investing in large-scale disinformation campaigns. It is extremely difficult for a new streamer to garner large numbers of concurrent viewers; it takes time to grow an audience on Twitch. Most Twitch content is also long-form and ephemeral. Since this means that most content is gone the moment it is created, it is not shared and does not go viral in the same way that it does on other UGC video and social media services, where videos are uploaded and can be viewed

and shared by users on demand. (ii) Our targeted policy only applies to those who persistently share harmful misinformation. Due to the long-form nature of Twitch’s content, we are focused on a streamer’s aggregated content rather than a specific, isolated statement within a longer piece of content. (iii) When we launched our Harmful Misinformation Actor policy, we took swift action against accounts that posed a threat to our community. We believe enforcement of our policy—particularly upon its adoption in H1 2022—has been an effective deterrent to harmful misinformation actors; we have not seen large numbers of them attempt to join our service.

Even if someone is not a Harmful Misinformation Actor, Twitch enforces on misinformation that targets specific communities under our Hateful Conduct & Harassment policies, and we take action on content that encourages others to engage in physically harmful behaviour under our Self-Destructive Behaviour policy. More information on enforcements under these policies can be found in [Twitch’s Safety Transparency Report](#).

Outcome 1b: Inform users about what content is targeted

Twitch’s Harmful Misinformation Actor policy is outlined in our [Community Guidelines](#) (CGs). Our aim is to have CGs that are clear and easy to follow but also thorough (with examples of prohibited behavior) to help Twitch users understand the boundaries we have set so they can feel confident expressing themselves within those boundaries.

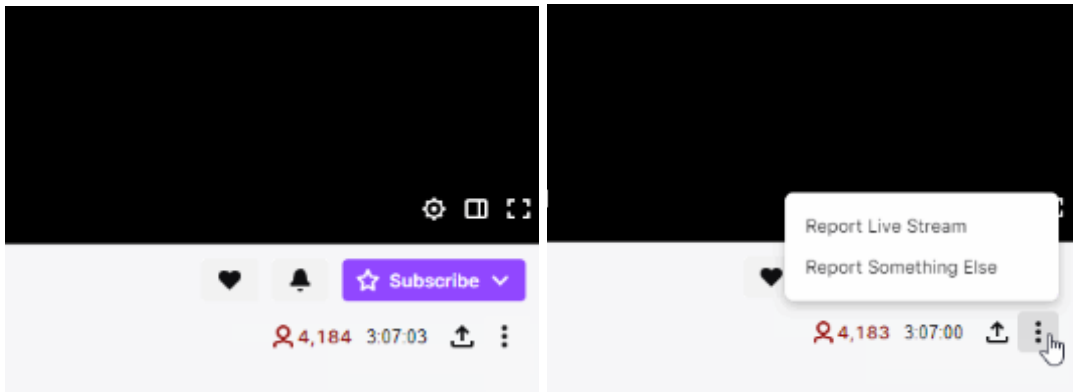
When we launched our Harmful Misinformation Actor policy in March 2022, we also published a [blog post](#) to flag the change for our community and provide more context on the policy.

Additionally, when a user violates our Harmful Misinformation Actor Policy—or any of our policies—they receive a detailed email notification. The notification includes the action taken, whether a suspension is permanent or temporary, the reason for the suspension, examples of violating content, a link to the Community Guidelines to learn more about the policy, where the violation occurred, and a link to the Appeals Portal if they disagree with the decision.

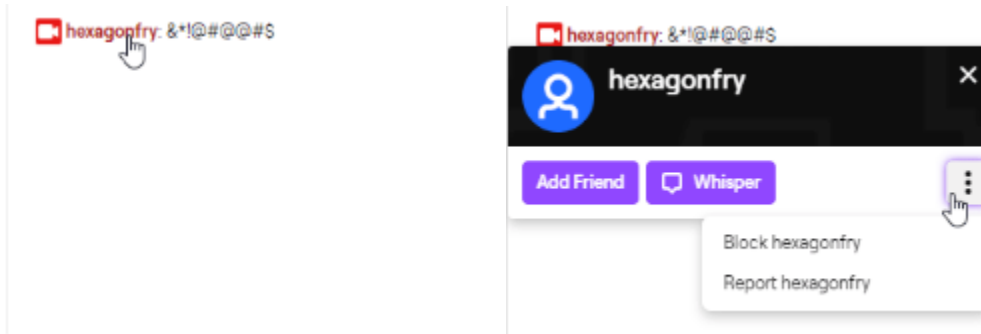
Outcome 1c: Users can easily report offending content

Users are able to report potential harmful misinformation actors with an easy, direct reporting mechanism. Users can submit a report by clicking on the three vertical dots icon, which is shown in the bottom right below the video player on the channel page; on the bottom right of a clip, highlight or past broadcast; or on the bottom right when you click on a username to report. An option to “Report Live Stream” or “Report [username]” will appear for the user to enter our reporting interface.

Reporting a Channel



Reporting a User in Chat



The interface then prompts the user to select the most relevant category for the violation, which in this case is “Misinformation.” Alternatively, a user can search for the appropriate reporting reason.

Reporting Interface

Report Chat Messages ×

Search for a reason

- Ban Evasion
- Account Ban Evasion
- Aiding Ban Evasion
- Bullying or Harassment
- Advocating Harassment
- Coordinating Harassment
- Malicious Pranks
- Revealing Personal Information
- Targeted Abuse

Back Next

Users can also submit a report to the specialised off-service investigations team through the team's email alias OSIT@twitch.tv.

Learn more about [how to file a report](#) on Twitch.

Outcome 1d: Information about reported content available

Twice-a-year, we publish a report outlining how we enforce our Community Guidelines, including our Harmful Misinformation Actor policy. This report is publicly available on our [website](#). We also provide a publicly-available [annual transparency report](#) under the EU Code of Practice on Disinformation.

Outcome 1e: Information about recommender engines

Twitch has published a [detailed summary of our various recommendation systems](#), outlining the main parameters used by each system. The page also provides information for how users can influence our recommendations and control what they see on Twitch.

When browsing, viewers can sort by 'Recommended for You' or by other channel attributes. Users can also customise their recommendations on Twitch by letting Twitch know if they are "not interested" in a streamer or content category that is recommended to them. At any time, users can navigate to their settings page and review what they have marked as "not interested" and then edit those selections. Users can learn more about this feature by visiting the [Help Article](#).

Objective 2: Disrupt advertising and monetisation incentives for disinformation.

Actors that systematically provide harmful misinformation are prohibited from the service, and are therefore not eligible for monetization. Additionally, [Twitch's ads policy](#) prohibits ads that contain deceptive, false, or misleading content as well as political content, such as campaigns for or against a politician, political party or related to an election, and/or content related to issues of public debate.

Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms.

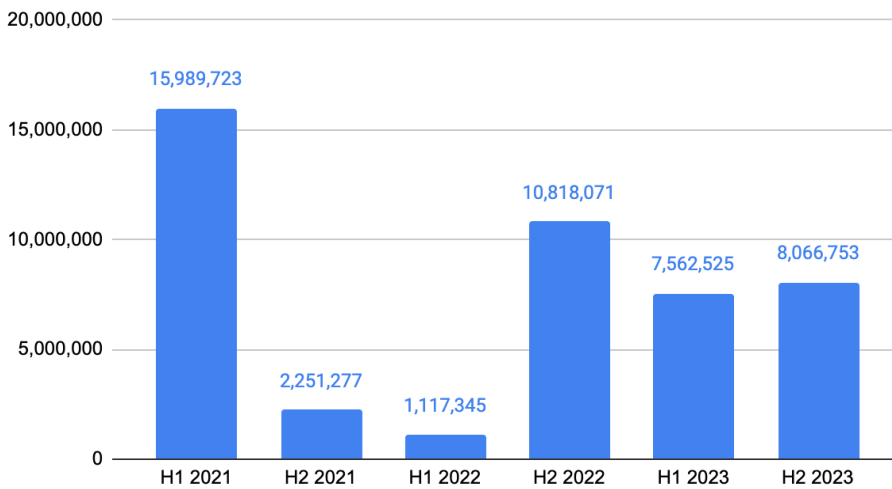
Twitch's [policies state](#) that "Any content or activity that disrupts, interrupts, harms, or otherwise violates the integrity of Twitch services or another user's experience or devices is prohibited." This includes the creation of inauthentic and malicious bots, impersonation, engaging in viewership tampering (such as artificially inflating follow or live viewer stats), and selling or sharing user accounts, services, or features.

We use historical enforcement data to proactively identify patterns associated with bots and spammers. Depending on the level of confidence, we can take several actions against a suspected bot account, including requesting that the account verify a mobile phone, auto-reporting the account to be reviewed by our operations team, and adding client-side friction that increases the cost of automation.

Most cases of impersonation on Twitch are phishing attempts, where a fraudulent channel is trying to get a user to click on a malicious link. We scan the text on our channel pages for these malicious URLs and then report the channel for review by our operations team. We also actively monitor channels for viewership tampering, using a combination of handcrafted filters based on ASN and IP reputation, as well as a machine learning model based on past examples.

In H2 2023, we issued 8.1M account enforcements for spam, scams, and fraud globally; 11,207 of these were for accounts based in Australia. Spam can be both automated (published by bots or scripts) or coordinated (when an actor uses multiple accounts to spread deceptive content). Due to its automated and coordinated nature, spam is generally Twitch's largest category of enforcement and we often see significant fluctuations in enforcement between reporting periods. This is consistent with a general trend in the industry.

Spam, Scams & Fraud Global Enforcements



Objective 4: Empower consumers to make better informed choices of digital content.

Twitch mitigates the risk that users are exposed to harmful misinformation on the site through the measures discussed previously. We are also committed to providing users with information about how our recommendation systems work and options to customise their recommendations as discussed under Outcome 1e above.

Twitch has also invested in a media literacy campaign to empower users to think critically about what information they consume. Twitch collaborated with media literacy expert MediaWise to develop an array of educational materials that teach Twitch streamers and viewers how to better identify, and avoid spreading, misinformation and disinformation online. These materials are hosted on the [Twitch Safety Center](#).

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.

Twitch remains open to supporting independent research if approached. At this time, Twitch does not directly support any third-party research.

Objective 7: Signatories will publicise the measures they take to combat Disinformation.

In Outcome 1b and 1e of this report, we provide details—and links to the corresponding materials—regarding publicly available information on the measures we take to combat misinformation.

Concluding remarks

As a signatory to the Australian Voluntary Code of Practice on Disinformation and Misinformation, Twitch is committed to combating misinformation on our service in an effective yet targeted manner that balances freedom of expression with keeping our communities safe. We recognize that harmful misinformation, and its prevalence on our platform, may evolve and we will continue to evaluate and adapt the measures we have put in place to protect our users and the integrity of our service.