# Microsoft report on Australian Voluntary Code of Practice on Disinformation and Misinformation

## Introduction

Microsoft is pleased to file this report on our commitments under the Australian Voluntary Code of Practice on Disinformation and Misinformation.

At Microsoft, we are committed to instilling trust and security across our products and services, and across the broader web. We also recognise that fighting disinformation is a key element to creating a trustworthy and safe online environment. Microsoft has a long history of working closely with governments, industry, civil society organisations, academics, and other stakeholders to ensure the integrity and security of our services and the online space more generally.

In addition to company-wide initiatives broadly aimed at promoting online trust and integrity, Microsoft also is working specifically to address the challenge of disinformation.

Our initiatives are not one size fits all, and as the code states, "…the types of user behaviours, content and harms that this code seeks to address will vary greatly in incidence and impact amongst the diverse range of services and products offered by different digital platforms."[1] In Microsoft's case, the majority of our services are used by enterprise customers or by individuals acting in a professional capacity.

This report details Microsoft's efforts to combat disinformation on our consumer-facing services, which we have detailed under each of the code's objectives/outcomes. **Appendix 1** includes a table that summarises each section of the code, and the relevant Microsoft initiative or policy.

## 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.*

[1] Australian Code of Practice on Disinformation and Misinformation, Preamble, p. 3.

- *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.*

Microsoft also has a number of programs in place to proactively combat disinformation on our services, as well technologies to help consumers verify the accuracy of online content.

### Video Authenticator

In September 2020, Microsoft announced a new technology, Video Authenticator, to address 'deep-fakes', or synthetic media, which are photos, videos or audio files manipulated by artificial intelligence (AI) in hard-to-detect ways.

Video Authenticator can analyse a still photo or video to provide a percentage chance, or confidence score, that the media is artificially manipulated. In the case of a video, it can provide this percentage in real-time on each frame as the video plays. It works by detecting the blending boundary of the deepfake and subtle fading or greyscale elements that might not be detectable by the human eye.

Microsoft expects that methods for generating synthetic media will continue to grow in sophistication. As all AI detection methods have rates of failure, platforms must understand and be ready to respond to deepfakes that slip through detection methods. Thus, in the longer term, there will be a need for stronger methods for maintaining and certifying the authenticity of news articles and other media. There are few tools today to help assure readers that the media they are seeing online came from a trusted source and that it was not altered.

### Coalition for Content Provenance and Authenticity

In February 2021 Microsoft was announced as a founding member of the Coalition for Content Provenance and Authenticity (C2PA).

The Coalition aims to address the prevalence of disinformation, misinformation and online content fraud through developing technical standards for certifying the source and history or provenance of media content.  The founding members of C2PA are Adobe, Arm, BBC, Intel, Microsoft and Truepic.

Although these standards are still in development, member organisations are working together to develop content provenance specifications for common asset types and formats to enable publishers, creators and consumers to trace the origin and evolution of a piece of media, including images, videos, audio and documents. These technical specifications will include defining what information is associated with each type of asset, how that information is presented and stored, and how evidence of tampering can be identified.

### Defending Democracy Program

In 2018 Microsoft launched the Defending Democracy Program, an innovative effort to protect democratic institutions and processes from hacking, to explore technological solutions to protect electoral processes, and to defend against disinformation. The Defending Democracy Program was created in response to the increasing threat posed to democratic processes by cyber-enabled interference, including attempts by nation-states to target and exploit key building blocks of our democratic system, as well as the manipulation of social media platforms to sow politically tinged misinformation.

The Defending Democracy Program also leverages Microsoft's role as a business and software provider to increase our clients' ability to counter outside efforts to compromise their security infrastructure. In August 2018, the Defending Democracy Program launched Microsoft AccountGuard, a free service for eligible Outlook 365 customers involved in elections, such as political parties and campaigns, political vendors, and think tanks. AccountGuard provides notifications about cyber threats, guidance on best practices for dealing with the unique problems faced by politically oriented organisations, and access to security features typically offered only to large corporate and government account customers. Currently, AccountGuard is available in 31 countries, including in Australia. These and other solutions that we offer, such as Microsoft ElectionGuard and Microsoft 365 for Campaigns, promote election integrity and campaign security. While AccountGuard and ElectionGuard are not targeted at disinformation specifically, they can help reduce its spread by creating obstacles to hacking by bad actors.

*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.*

The Microsoft Services Agreement for consumer products outlines the types of behaviours and types of content that are prohibited when using Microsoft products and services.

The Code of Conduct included in the Microsoft Services Agreement prohibits a number of behaviours that may relate to the spread of disinformation and misinformation. The rules include:

- Don't send spam or engage in phishing
- Don't engage in activity that is fraudulent, false or misleading (e.g., asking for money under false pretenses, impersonating someone else, manipulating the Services to increase play count, or affect rankings, ratings, or comments)
- Don't engage in activity that is harmful to you, the Services or others.

Additionally, the Microsoft Services Agreement provides for enforcement measures, which may include the removal of services or the closure of a Microsoft account. Other enforcement measures include the removal of content or refusal to publish content.

LinkedIn also has similar terms of use in place and enforces a blanket prohibition on automated activity on the platform. LinkedIn's Professional Community Policies also include guidelines that prohibit impersonation, misrepresentation and harmful or violent content on the platform.

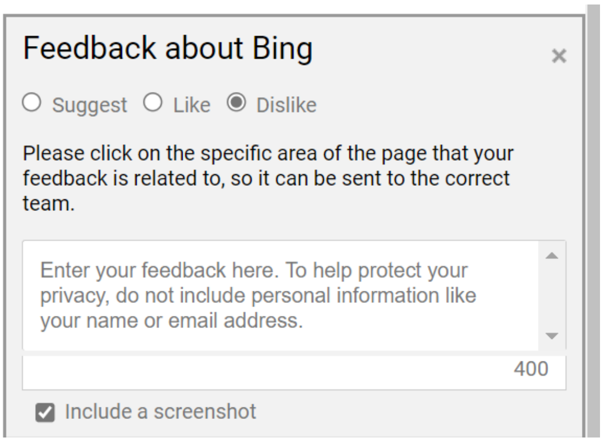
*Outcome 1c: Users can report content and behaviours to Signatories that violates their policies under 5.10 through publicly available and accessible reporting tools.*

- *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.*

- *In implementing the commitment in 5.11 signatories recognise that the terms Disinformation and Misinformation may be unfamiliar to users and thus policies and procedures aimed at achieving this outcome may specify how users may report a range of impermissible content and behaviours on digital platforms.*

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

For instance, on Bing, users can use the feedback tool (see below) to select a specific search result to flag. The user can include a screenshot of the specific search result, as well as a description of the issue. Additionally, Bing users can also use the Report a Concern page to directly report other types of content, as well as concerns about misinformation and disinformation through the 'other concern' option.



Feedback about Bing

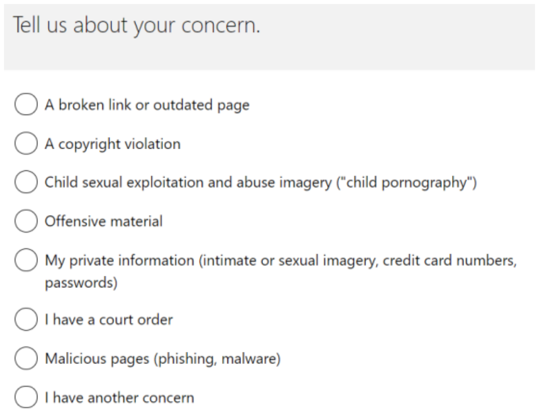
○ Suggest ○ Like ● Dislike

Please click on the specific area of the page that your feedback is related to, so it can be sent to the correct team.

Enter your feedback here. To help protect your privacy, do not include personal information like your name or email address.

400

☑ Include a screenshot



Tell us about your concern.

○ A broken link or outdated page

○ A copyright violation

○ Child sexual exploitation and abuse imagery ("child pornography")

○ Offensive material

○ My private information (intimate or sexual imagery, credit card numbers, passwords)

○ I have a court order

○ Malicious pages (phishing, malware)

○ I have another concern

Similarly on LinkedIn, a user can report a piece of content by reporting a specific post. The 'More' icon on the right side of the post/comment enables a user to 'Report this post.' The user can then select the option that best describes their concerns with the post/comment from the 'Why are you reporting this?' pop-up.

Additional information is also available on LinkedIn's website about reporting Inappropriate Content, Messages, or Safety Concerns.

*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.*

Microsoft regularly publishes information about the detection and removal of content that violates our user policies as outlined below.

**LinkedIn Community Report**

The LinkedIn Community Report describes actions we take on content that violates our Professional Community Policies and User Agreement. The latest version of this report covers the six-month period between January 1 and June 30, 2020.

The report covers the detection of fake accounts, spam and scams, content violations and copyright infringements. In the January to June 2020 period:

- 33.7 million fake accounts were stopped at registration;
- 98% of fake accounts were detected automatically;
- 79.3 million scams/spam were proactively detected and removed by LinkedIn; and
- 22, 846 pieces of misinformation content were removed by LinkedIn.

**Microsoft Digital Safety Content Report**

Our digital safety content report provides information on actions that Microsoft has taken in relation to child sexual exploitation and abuse imagery (CSEAI), terrorist and violent extremist content (TVEC), as well as non-consensual intimate imagery (NCII).

This report provides information on the content and accounts, upon which we have taken action through removal and suspensions.

This report is inclusive of Microsoft consumer products and services including (but not limited to) OneDrive, Outlook, Skype, Xbox and Bing. This report does not include data representing LinkedIn or Github which have their own transparency reports.

**Microsoft Law Enforcement Requests Report**

Twice a year we publish the number of legal demands for customer data that we receive from law enforcement agencies around the world. While this report only covers law enforcement requests, Microsoft follows the same principles for responding to government requests for all customer data.

## Objective 2: Disrupt advertising and monetisation incentives for Disinformation.

*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.*
- *Signatories recognise that all parties involved in the buying and selling of online advertising and the provision of advertising related services need to work together to improve transparency across the online advertising ecosystem and thereby to effectively scrutinise, control and limit the placement of advertising on accounts and websites that propagate Disinformation.*

Microsoft's company-wide commitment to disrupting the economics of disinformation is well-illustrated by Microsoft Advertising (formerly Bing Ads), our proprietary advertising platform, which serves the vast majority of ads displayed on Bing, and provides advertising to most other Microsoft services that display ads, as well as many third-party services. Microsoft Advertising works both with advertisers, who provide it with advertising content, and publishers, such as Bing, who display these advertisements on their services. As detailed below, 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.

### A. Advertisers

With regard to advertisers, 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 organizations. Microsoft Advertising also has a set of Relevance and Quality Policies to 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).

Microsoft Advertising employs dedicated operational support and engineering resources to enforce these policies, combining automated and manual enforcement methods to prevent or take down advertisements that violate its policies. Every ad loaded into the Microsoft Advertising system is subject to these enforcement methods, which leverage machine-learning techniques, automated screening, the expertise of its operations team, and dedicated user safety experts. In addition, Microsoft Advertising conducts a manual review of all advertisements flagged to its customer support team and removes advertisements that violate its policies.

Across all markets in 2020, Microsoft suspended nearly 300,000 accounts from the Microsoft Advertising platform, up 30% from 2019. We also removed 1.6 billion bad ads while 270,000 sites were also removed from our system. In response to COVID we put strict measures in place to restrict advertising for products such as covid medicines. We rejected nearly 21 million ads in accordance with our sensitive advertising policy.

### B. Publishers

With regard to publishers, Microsoft Advertising utilises a distinct set of measures designed to avoid the display of advertising on—and thus disrupt the flow of advertising revenue to—sites involved in spreading disinformation.

First, 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.

These publishers also benefit from the set of measures identified above that Microsoft Advertising takes with regard to advertisers, which ensures that these partners receive high-integrity, non-deceptive ads from the Microsoft Advertising platform.

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, sensitive political 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 organizations). Publishers are required to maintain a list of prohibited terms and provide us 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).

Microsoft Advertising reviews publisher properties and domains for policy compliance, including compliance with restrictions on prohibited content. In this review, Microsoft Advertising also considers feedback from its advertisers to help ensure a safe environment for the delivery of their advertisements and maintains a review process to investigate related advertiser complaints. Publishers are promptly notified of properties or domains that violate Microsoft Advertising's policies; such properties and domains are not approved by Microsoft for live ad traffic. If a property or domain is already live, and later found in violation of Microsoft Advertising's policies, it is removed from the network until the publisher remedies the issue.

## 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.*

- *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.*

- *To allow for the expectations of some users and digital platforms about the protection of privacy, measures developed and implemented in accordance with this commitment should not preclude the creation of pseudonymous and anonymous accounts.*

Microsoft understands that protecting the integrity of our services against automated actors and false accounts is essential to maintaining a safe and trusted online environment for our customers.

We list below several measures that Microsoft takes to maintain the integrity of our services against threats from bots and other false account activity in relation to Bing, Microsoft Advertising, and LinkedIn:

**1. Bing**

Because users cannot post content to Bing directly, it is not vulnerable to certain types of harmful bot and false account activity that afflict many other types of online services, such as social media services. Instead, Bing search results draw on an index of the web created by Microsoft itself by crawling billions of URLs, which are ranked using proprietary algorithms, thereby limiting substantially the possibility for many common types of abuse.

That does not mean that bots do not pose a threat to the service, however. Indeed, Bing takes significant efforts to ward off other types of bot activity. For instance, Bing observes attempts to abuse its search platform by bots seeking to influence its ranking system (e.g., by engaging in "click fraud" or harnessing "link farms" to make it seem as though a site is more popular than it actually is). This ranking process is the facet of Bing that could be vulnerable to manipulation by bots, as bad actors attempt to game the myriad heuristics Bing Search uses to determine how authoritative a specific URL is, and thus how high in the search results pages it should be positioned.

To combat this type of abuse, Bing Search implements a sophisticated ranking process that is increasingly focused not only on ensuring that users always see the most relevant results for their search query on the first page, but also that such results are "high authority" (unless a user's query clearly intends to find low authority content). Factors that impact a page's rank on Bing include: relevance; quality and credibility; user engagement; and location. More information on how Bing ranks content is published in the Bing Webmaster Guidelines.

As background, Bing determines the "authority" of a page using a variety of heuristic signals of the relative relevance of a page to a user's search keywords, including, for example: reputation (sometimes gauged by the number and quality of links to that site by third party sites); the level of discourse on the site (i.e., sites for which the purpose of the posted content is solely to denigrate without adding substantively to the discourse will be seen as lower authority); the level of distortion (i.e., the degree to which the site differentiates opinion from fact); degree of origination (i.e., whether the site reports first-hand information, as opposed to republishing content from elsewhere); whether the site is spam (i.e., whether it attempts to manipulate its ranking); and transparency of authorship and sourcing information.

Disinformation, however, may at times appear in both organic and paid search results, and we take active steps to counter it, as described above. It is worth emphasising, however, that 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.

In the case of material that is known (or suspected) to be false, Bing generally does not remove such content from search results as a policy matter. Blocking content in organic search results can raise significant fundamental rights concerns relating to freedom of expression and the freedom to receive and impart information. While Bing strives to rank its organic search results so that trusted, authoritative news and information appear first, and provides tools that help Bing users evaluate the trustworthiness of certain sites (as described in more detail below), we also believe that enabling users to find all types of information through a search engine can provide important public benefits.

For example, even with known fake articles, a researcher or journalist may actually want to find such unreliable information for legitimate reasons. As such, Bing removes content from search results only in limited cases, such as where content is illegal pursuant to local law (e.g., copyright infringement, defamation, or the EU's Right to be Forgotten), or where such content is universally abhorrent (e.g., child sexual exploitation and abuse imagery or nonconsensual pornography).

### 2. Microsoft Advertising

Microsoft Advertising, as an online ad network, maintains the integrity of its services in part though the same measures it takes to scrutinise ad placements. Microsoft Advertising also 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. Microsoft Advertising has several teams of security engineers, support agents, and traffic quality professionals dedicated to continually developing and improving this traffic filtration and network monitoring system. Microsoft Advertising's support teams work closely with its advertisers to review complaints around suspicious online activity, and they work across internal teams to verify data accuracy and integrity.

### 3. LinkedIn

As a real identity professional network whose value lies in its use by individuals for professional purposes, LinkedIn acts vigilantly to maintain the integrity of all accounts and to ward off bot and false account activity. LinkedIn vigorously 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." LinkedIn employs a dedicated Anti-Abuse team to create the tools to enforce this prohibition. LinkedIn's automated systems detect and block automated activity, impose hard limits on certain categories of activity commonly engaged in by bad actors, and detect whether members have installed known prohibited automation software; LinkedIn also conducts manual investigation and restriction of accounts engaged in automated activity on the LinkedIn platform. LinkedIn is also partnering with the broader Microsoft organization to develop technological solutions for detecting "deepfakes"—i.e., synthetic imitations of real individuals created using artificial intelligence and other advanced technological means.

## 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.

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

Microsoft is committed to helping our customers make informed decisions about what 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.

### 1. News

Our Microsoft News service provides users with quality journalism by partnering with over 1000 reputable news sources worldwide, each of which is vetted by Microsoft News' team. In addition, the Microsoft News service may include news search results from Bing News. These search results must meet the Bing News content standards found here https://pubhub.bing.com/Home/Help.

Microsoft also helps its customers evaluate the quality of the news they encounter on the Internet through its partnership with NewsGuard Technologies. Led by respected journalists and entrepreneurs Steven Brill and Gordon Crovitz, NewsGuard's analysts review online news sites across a series of nine journalistic integrity criteria, such as whether the site regularly publishes false content, reveals conflicts of interest, discloses financing, and publicly corrects reporting errors.

NewsGuard then compiles their findings into a "Nutrition Label" and corresponding Red/Green Reliability Rating, which users can view to better understand the reliability of the news they consume.

Microsoft has partnered with NewsGuard to provide a free plug-in for the Microsoft Edge web browser (also available for other browsers including Chrome and Firefox), as well as an opt-in news rating feature for the Edge mobile application on both iOS and Android. This empowers Edge users to benefit from the comprehensive analysis done by NewsGuard and to better identify the most reliable news and information sites.

### 2. Search results

Bing is taking two different sets of measures to enable users to identify and discover trustworthy content when engaging in online search. First, Bing has created tools to help users independently evaluate the legitimacy of content they find online through Bing search results. Second, Bing is continually refining its algorithms to ensure it returns the highest authority results possible to its users.

As an example of the first type of measure, Bing offers an "Intelligent Search" feature in response to certain user queries that demonstrate an intent by a user to learn all valid answers to a question. Identifying questions with multiple valid answers involves several techniques, including sentiment analysis, which identifies the opinions expressed in a piece of text (i.e., positive, negative or neutral). Through this feature, Bing will summarise the various potentially valid answers to a user query and display them all in a carousel to give the user a balanced overview.

Another example of how Bing is working to assist users in determining the reliability of content displayed in Bing search results is its "Fact Check" feature. Introduced in September 2017, Fact Check helps users find fact checking information on news displayed within Bing search results.

Bing also empowers consumers by continuously improving its ability to accurately and instantaneously place sites along a high / low authority continuum. Research indicates that websites that promulgate disinformation tend to meet Bing's internal classification standards for "low authority" sites. As Bing gets better and better at differentiating high from low authority material, consumers will be exposed to fewer sites of low reliability (including sites disseminating disinformation) when searching for information and content online.

### 3. User-generated content

Microsoft is also working to help our customers identify trustworthy user-generated content on our services. LinkedIn, for example, has implemented multiple programs to advance this goal. LinkedIn provides members with "best practices" guidelines for sharing quality content on its platform. LinkedIn also has an Influencer Program of over 500 handpicked leaders in their respective industries who create and share content on LinkedIn. Similarly, LinkedIn helps members find the news and conversations that are most relevant to them by creating curated interest-based feeds about the most important developing stories in a member's industry. By creating a place on LinkedIn where editors spotlight the best conversations from its members on a news story, with diverse voices curated by a team of dozens of in-house journalists, LinkedIn is able to reward members who start those conversations by allowing their contributions to reach a broader audience and encourage other diverse voices to participate.

### 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.*

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

- *Signatories may also, as a matter of policy, choose not to target advertisements based on the inferred political affiliations of a user.*

On April 15, 2019, Microsoft Advertising updated its Advertising Policies to prohibit 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. Furthermore, Microsoft prohibits political advertising across Microsoft media properties and platforms pursuant to our Creative Acceptance Policy and Native Creative Acceptance Policy.

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" is also prohibited.

To the extent Microsoft uses services other than Microsoft Advertising to display ads on their sites, those services likewise enforce prohibitions on the use of political advertisements. For instance, LinkedIn uses automated and manual review to enforce its Advertising Policies, which provide that "Political ads are prohibited, including ads advocating for or against a particular candidate or ballot proposition, or otherwise intended to influence an election outcome." 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.*

- *Signatories commit to support and encourage good faith independent efforts to research Disinformation and Misinformation both online and offline. Good faith research includes research that is conducted in accordance with the ethics policies of an accredited Australian University provided such policies require that data collected by the researcher is used solely for research purposes and is stored securely on a university IT system or any research which is conducted in accordance with the prior written agreement of the digital platform.*

- *Signatories commit not to prohibit or discourage good faith research into Disinformation and Misinformation on their platforms.*

- *Relevant Signatories commit to convene an annual event to foster discussions regarding Disinformation within academia and Civil Society.*

The research community has a critical role to play in helping to develop new methods and insights on how best to stamp out disinformation online. Accordingly, Microsoft is working at a company-wide level to ensure that the research community has the tools and resources it needs to play its part in helping to combat disinformation.

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

**Technology | Academics | Policy (TAP).** Microsoft provides administrative and financial support for TAP, a forum for leading academics to share articles and research that focus on the impact of technological change on society, and to share ideas with each other and the broader online community.

**Woodrow Wilson Disinformation Collaboration.** Through an ongoing partnership with Jake Shapiro at the Princeton University Woodrow Wilson School of Public and International Affairs on a project titled "Trends in Online Foreign Influence Operations," Microsoft's

Defending Democracy Program has explored topics such as the link between malicious content on social media (where many bots and trolls operate) and the wider news ecosystem. Among other key findings, the project concluded that 24 different countries were targeted by foreign influence efforts between 2013 and 2018. Working with Princeton's Center for Information Technology Policy, Microsoft has also supported the development of new tools dedicated to detecting foreign information operations and influence programs, particularly within the United States, the largest single target of foreign disinformation campaigns. The partnership draws on the expertise of both Microsoft Research and Princeton's technical and political science teams.

**Oxford Internet Institute.** Microsoft is partnering with and funding the Oxford Internet Institute at Oxford University for a two-part research project. The Institute is developing an 'International Election Observatory,' which will deploy a dedicated team to track propaganda usage during elections (the Institute has already observed elections in Sweden, Brazil, the United States, the European Union, and India). The Observatory will then use tools developed with the support of the National Science Foundation and European Research Council to develop and share their research results with policymakers, the media, and the public. Microsoft also provided financial support for the launch of the Oxford Technology & Elections Commission, which in August 2019 released a Report on Anti-Disinformation Initiatives, investigating fake news landscapes around the world.

**Trust Project.** Bing News has joined the Trust Project, a consortium of top news and digital companies led by journalist Sally Lehrman at Santa Clara University's Markkula Center for Applied Ethics. The Trust Project brings together, among others, Google, Facebook, and top news organizations (including the Washington Post, New York Times, BBC, and Globe and Mail) to identify new ways to make it easier for the public to identify the quality of news. By leveraging information from the Trust Project, Bing can determine how to better inform users about the reliability of news articles and help deliver on its promise of providing transparent, balanced and trustworthy search results within Bing News.

**Microsoft Research Lab.** In Microsoft Research's lab in New York City, Microsoft brings together social scientists and humanists from the fields of anthropology, communication, economics, information, law, media studies, women's studies, science & technology studies, and sociology to provide insight into how social media is reconfiguring society. Microsoft Research also maintains the Social Media Collective, a network of social science and humanistic researchers including Nancy Baym, Danah Boyd, Kate Crawford, Tarleton Gillespie, and Mary Gray. The primary purpose of the Social Media Collective is to enhance understanding of the social and cultural dynamics that underpin social media technologies, which ultimately should prove helpful to combating disinformation on social media services.

**AI and Ethics in Engineering and Research Committee ("AETHER")**. Microsoft's AETHER Committee—a group of senior leaders from across the company that formulates internal policies on the responsible use of artificial intelligence—has created the Focus Team on Authenticated Media Provenance ("AMP") in collaboration with leading media organizations. AMP uses a combination of smart-caching of audio and video to apply a "fragile watermark" to media. This fragile watermark then serves as a visual cue of the media's authenticity, persisting notwithstanding basic editing (e.g., cropping, relaying, or 17 retransmission of video), but not in the face of substantial editing (e.g., face swapping or audio dubbing).

In addition to these and other initiatives, Microsoft continues to engage in pioneering research of our own in the disinformation space. For example, Microsoft is leading the research into the phenomenon of data voids. Data voids are search terms for which the available relevant data is limited, non-existent, or deeply problematic; when a user queries for a data void, a search engine's prioritization of high authority results will not prevent the user from receiving disinformation in response to their search. Microsoft is leading the development of strategies for preventing the exploitation of data voids by individuals with ideological, economic, or political agendas.

Microsoft is also working to better understand and address the emerging threat posed by the use of artificial intelligence tools to develop malicious synthetic media (i.e., "deepfakes"). Although synthetic media tools have many benign uses and potential benefits, deepfakes can be used to damage reputations, fabricate evidence, and undermine trust in our democratic institutions. To help guard against this challenge, Microsoft is engaging with partners in academia, civil society, and the technology industry through forums like the Partnership on AI to identify possible countermeasures against deepfakes. Microsoft has also supported interdisciplinary research within the company in a number of disinformation-related areas.

## 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.*

- *All Signatories will make and publish the annual report information in section 7*
- *In addition, Signatories will publish additional information detailing their progress in relation to Objective 1 and any additional commitments they have made under this Code.*

In addition to our reporting under this code, Microsoft releases information about our initiatives to combat disinformation.

Microsoft's On the Issues blog contains company announcements about technology policy issues, including those relating to disinformation and misinformation. More information on our Microsoft's latest announcements—our video authenticator Video Authenticator technology and the Coalition for Content Provenance and Authenticity (C2PA) —can be found on this blog.

## Conclusion

Microsoft is committed to playing our part in stamping out disinformation online. This report summarises some of the key initiatives Microsoft is taking both as a company and on a service-by-service basis to promote the goals of, and comply with the commitments in, the Australian Voluntary Code of Practice on Disinformation and Misinformation.

## APPENDIX 1: MICROSOFT RESPONSES TO CODE OBJECTIVES AND OUTCOMES

| Objective | Outcome | Links to Policies/Actions | Explanation |
|---|---|---|---|
| **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. | Video Authenticator<br>Coalition for Content Provenance and Authenticity (C2PA)<br>Defending Democracy Program | Microsoft has a number of programs in place to proactively combat disinformation on our services, as well technologies to help consumers verify the accuracy of online content. |
| | 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. | Microsoft Services Agreement | The Microsoft Services Agreement for consumer products outlines the types of behaviours and types of content that are prohibited when using Microsoft products and services. |
| | Outcome 1c: Users can report content and behaviours to Signatories that violates their policies under 5.10 through publicly available and accessible reporting tools. | Bing Report a Concern<br>LinkedIn - Report Inappropriate Content, Messages, or Safety Concerns. | Both Microsoft and LinkedIn have reporting mechanisms where users are able to flag problematic content. |
| | Outcome 1d: Users will be able to access general information about Signatories actions in response to reports made under 5.11. | LinkedIn Community Report<br>Microsoft digital safety content report | Microsoft regularly publishes information about the detection and removal of content that violates our user policies. |

| | | | |
|---|---|---|---|
| **Objective 2: Disrupt advertising and monetisation incentives for Disinformation.** | Outcome 2: Advertising and/or monetisation incentives for Disinformation are reduced | LinkedIn Professional Community Policies | LinkedIn's Professional Community Policies outline the activities that are acceptable on the service, and what is inappropriate and may be stopped by LinkedIn. |
| | | LinkedIn Advertising Policy and LinkedIn User Agreement | LinkedIn's advertising guidelines outlines the restrictions on content for advertising on the platform. LinkedIn's User Agreement is the contract which outlines the policies for use of its service. |
| | | Microsoft Services Agreement | The Microsoft Services Agreement outlines the terms and service and enforcement related to activity that is considered fraudulent, false or misleading. |
| | | Microsoft Advertising policies | Microsoft's advertising policies help advertisers learn  what is and is not allowed. |
| | | Creative Acceptance Policy and Native Creative Acceptance Policy | Third party ad platforms must abide by these guidelines for the display and native advertising served across all Microsoft publishers and markets (except for LinkedIn); These ad policies prohibit misleading messaging, content or images, including unsubstantiated claims or endorsements, have the potential to be interpreted as misleading using sensationalized text, as well as messaging/content that is not related to the landing page, or product/service being promoted. |

| | | | |
|---|---|---|---|
| **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. | LinkedIn's User Agreement | LinkedIn has and enforces a blanket prohibition on automated activity on the platform. |
| | | LinkedIn Professional Community Policies | LinkedIn's Professional Community Policies include guidelines that prevent impersonation and misrepresentation on the platform. |
| | | Bing News PubHub Guidelines for News Publishers | Bing News' PubHub guidelines for news publishers helps safeguard that the vertical's search and news results are as comprehensive, relevant, balanced and trustworthy as possible. This also includes verification policies for news sites. |

| | | | |
|---|---|---|---|
| **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. | Blog post on Microsoft's partnership with NewsGuard | Microsoft is partnering with NewsGuard to provide a free Edge Desktop browser plug-in and an integrated add-on to the Edge mobile apps on iOS and Android. The add-on provides more information to consumers about the quality of the news they consume by automatically flagging news sites with a Red/Green credibility and transparency rating.<br>The ratings are based on nine journalistic criteria—such as whether the site regularly publishes false content, reveals conflicts of interest, discloses financing, or publicly corrects reporting errors. Microsoft does not have any oversight or editorial control over NewsGuard's ratings. |
| | | Reporting of fake news on LinkedIn | LinkedIn prohibits posting content that is intentionally deceptive, including fake news, as outlined in its User Agreement and Professional Community Policies.  To help LinkedIn remain a trusted environment for professionals and keep false information off our platform, we have created a way for LinkedIn users (known as "members") to report or "flag" misleading news. LinkedIn members can also report misleading or false content through our Help Center.<br>Whenever a member reports abuse using that channel, a customer service representative will evaluate and take appropriate action. |

| | | | |
|---|---|---|---|
| | | Bing and Bing News Fact Checking Feature | Bing and Bing News' Fact Check feature brings more information to users about what they can trust. |
| | | Trust Project | Bing News is part of the Trust Project which is an international consortium of news organizations building standards of transparency and working with technology platforms to affirm and amplify journalism's commitment to transparency, accuracy, inclusion, and fairness so that the public can make informed news choices. |
| | | Microsoft ad settings for users | Microsoft enables users to tailor the ads they are seeing. |
| | | Microsoft privacy dashboard | Microsoft's policies for how and why it collects and use personal data. Users are able to access and manage their personal data through the Microsoft privacy dashboard. |
| **Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.** | | LinkedIn's advertising policy | Political ads are prohibited on LinkedIn, including ads advocating for or against a particular candidate or ballot proposition, or otherwise intended to influence an election outcome. |
| | | Microsoft Advertising policies | Microsoft Advertising prohibits ads for election related content, political parties and candidates, and ballot measures globally. |

| | | | |
|---|---|---|---|
| **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. | Microsoft blog post on defending against disinformation | Microsoft supports academic research into disinformation and participates in several partnerships globally. |
| **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. | Microsoft's On the Issues blog | In addition to our reporting under this code, Microsoft releases information about our initiatives to combat disinformation.<br>Microsoft's On the Issues blog contains company announcements about technology policy issues, including those relating to disinformation and misinformation. |