



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](#).

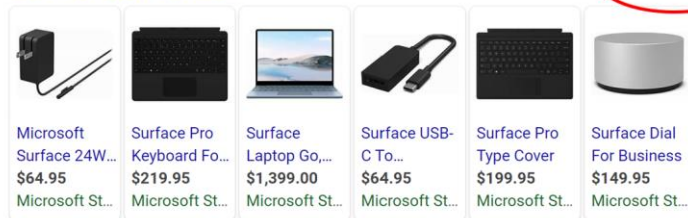
Surface Devices, Accessories - Microsoft Store

<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.

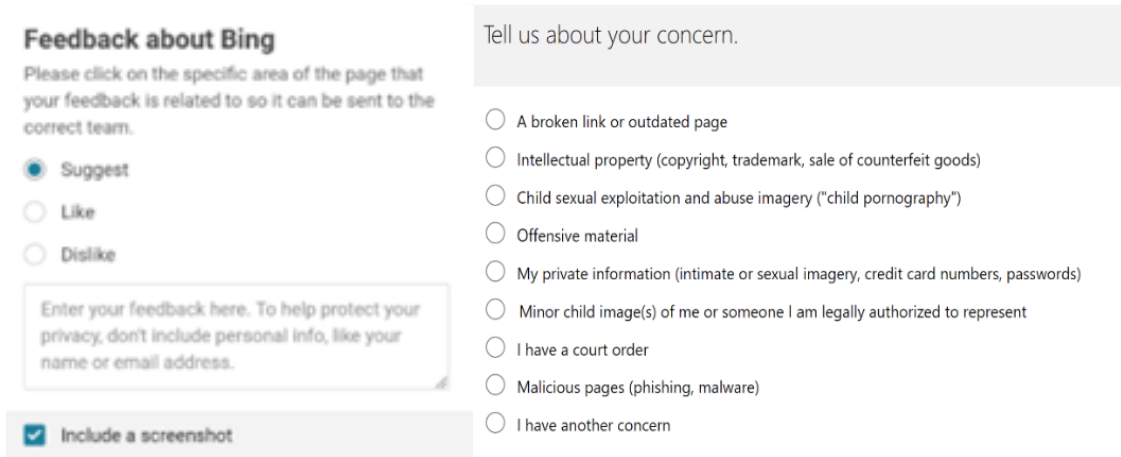
(O.1c) Bing Search

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.



Feedback about Bing

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

Suggest
 Like
 Dislike

Enter your feedback here. To help protect your privacy, don't include personal info, like your name or email address.

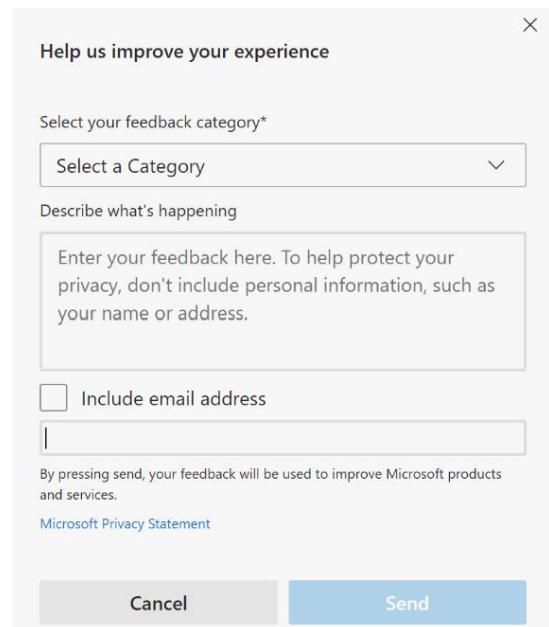
Include a screenshot

Tell us about your concern.

- A broken link or outdated page
- Intellectual property (copyright, trademark, sale of counterfeit goods)
- Child sexual exploitation and abuse imagery ("child pornography")
- Offensive material
- My private information (intimate or sexual imagery, credit card numbers, passwords)
- Minor child image(s) of me or someone I am legally authorized to represent
- I have a court order
- Malicious pages (phishing, malware)
- I have another concern

(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.



Help us improve your experience ×

Select your feedback category*

Select a Category ▾

Describe what's happening

Enter your feedback here. To help protect your privacy, don't include personal information, such as your name or address.

Include email address

By pressing send, your feedback will be used to improve Microsoft products and services.

[Microsoft Privacy Statement](#)

Cancel Send

(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

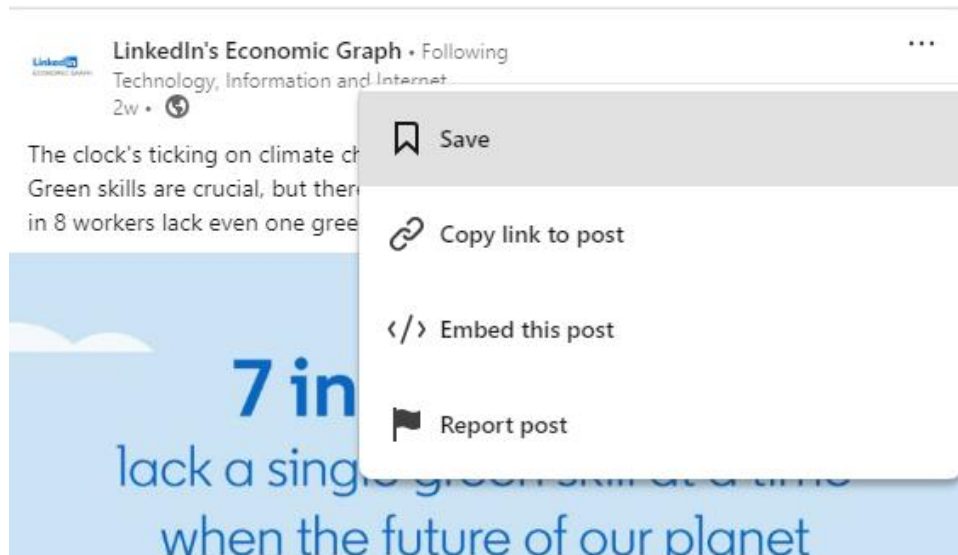
- I would like information, tips and offers about Microsoft Advertising. [Privacy Statement](#)

 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](#)

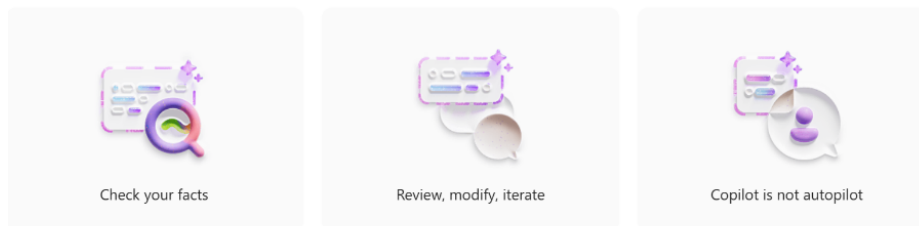
[The Copilot in Bing FAQs](#) and similar explanatory documents like [blog posts](#) and [Copilot in Bing: Our Approach to Responsible AI](#) also help educate users on the nature of AI-driven search experiences and the uses, safeguards, and limitations of this emerging technology.

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>
MS MARCO	<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.