Bing

Bing Systemic Risk Assessment Report

August 2024

Microsoft

Executive Summary 3
Introduction 5
How to read this document 7
Bing overview 8
  User base and market 8
  Core search 9
  Generative AI features 13
  Ancillary search features 16
  New Bing features 20
Bing's approach to protecting users and their fundamental rights online 23
  Bing's values and commitments 23
  Impact to risk profile and key risk considerations 25
  Service-wide mitigations 26
  Bing's approach to risk on core search 29
  Bing's approach to risk on generative AI features 33
  Bing's approach to risk on ancillary search features 40
  How Bing monitors effectiveness of risk mitigation efforts 44
Systemic Risk Assessment methodology 47
  Risk assessment design enhancements 47
  Risk assessment process 48
  Assessment tools 51
Risk assessment results 52
  Areas of Moderate Residual Risk 52
  Areas of Low Residual Risk 81
A**dditional insights** 139
  Changes to Bing's risk profile 139
  Progress on enhanced mitigations from Bing's last report 141
  Looking ahead to the coming year 143
Conclusion 145
Appendix I: Detailed Risk Assessment and Scoring Methodology 147
Appendix II: Catalog of Mitigations by Industry Best Practices 156
Appendix III: Abbreviated Terms 181

Executive Summary

In a technology-driven world, search engines are in many ways the gateway to the Internet and the primary way people find the content they are looking for amongst the trillions of ever-changing webpages available online. Search engines like Bing play a vital role in upholding the fundamental rights of free and open access to information and free expression. At the same time, Bing recognizes that other fundamental rights and social interests, such as privacy, safety, democratic processes, public health, and national security, are also vital to Bing's users and to a healthy society, and Bing must balance these interests and rights to maximize benefit to users while minimizing possible harms. Addressing content risks in a search service often requires a different approach than may be appropriate for other types of online services. Over-moderation of content in search could have a significant negative impact on the right to access information, freedom of expression, and media plurality. Therefore, Bing must carefully balance these competing fundamental rights and interests as it works to ensure its algorithms return the highest quality content relevant to a user's queries without unduly limiting their ability to access answers to the questions they ask.

Bing has implemented robust and multi-tiered risk mitigation measures to maintain this careful balance. These mitigations are anchored in the principles of:

- developing ***fair, reliable, safe, private, secure, inclusive, transparent,*** and ***accountable*** algorithmic systems;
- providing ***credible and authoritative results*** relevant to user queries;
- ***promoting free and open access to information*** within the bounds of the law and with respect for local law and other fundamental rights, such as privacy and public safety;
- taking steps to ***protect users from harmful and unexpected offensive content***; and
- ***being transparent*** with users about Bing's principles and practices, as well as Bing's decisions and actions.

Teams across Bing continuously measure the effectiveness of the risk mitigation measures implemented in alignment with these principles to adjust and refine them as threats and risks evolve. The results of this continuous monitoring indicate a low occurrence of systemic harms on Bing.

Bing is committed to continually enhancing user trust and safety and has conducted an assessment of the effectiveness, proportionality, and reasonableness of the implemented mitigations relative to potential systemic risks stemming from the use, misuse, or functioning of the service pursuant to Article 34 of the EU Digital Services Act. Bing has implemented several enhancements to its Systemic Risk Assessment process, including implementing a quantitative risk scoring methodology, conducting additional internal and external consultations, and expanding on specific risk scenarios and how they could theoretically manifest on various Bing features to support the evaluation of mitigation measures.

As part of this process, Bing has identified opportunities to further enhance the systemic risk mitigations currently in place. These opportunities include implementing:

- a more direct mapping of Article 34 enumerated systemic risks within existing risk monitoring processes;
- more refined user reporting capabilities across Bing enhanced search and generative AI features;
- expanded protections across enhanced search features to further address potential crisis events and safeguards for minor users;
- continued investment in classifiers and mechanisms for identifying evolving harms on the service, including malware and content related to sexual exploitation; and
- increased engagement with experts to improve understanding of how risk areas manifest or are experienced by Bing users to consider additional or refined mitigation measures.

Bing will work to implement these enhanced mitigations and will continue to monitor evolving risks on the service to enhance and refine measures as needed to maintain Bing's overall low risk profile.

Introduction

This Systemic Risk Assessment Report (the Report) is responsive to the obligations listed in Article 42.4 (a) and (b) of regulation (EU) 2022/2065, known as the Digital Services Act (DSA), in setting out the results of the Systemic Risk Assessment, which was conducted pursuant to Article 34, and the specific mitigation measures put in place pursuant to Article 35(1).

The Report addresses the Bing service, which was designated by the European Commission on 25 April 2023 as a Very Large Online Search Engine (VLOSE). The Bing service includes the services offered on Bing.com, Bing mobile apps, and relevant features, as described in the section Bing Overview. This includes core search features; enhanced search features, including search verticals such as images, shopping, maps, news, travel, and real estate, and advertisements in Bing; and the generative AI-enhanced features, Copilot in Bing, and Image Creator from Bing.

The report describes the assessment of systemic risks in the Bing VLOSE impacting the EU for the period of August 2023-July 2024.

The report discusses key measures implemented to mitigate these risks; the evaluation of their reasonableness, effectiveness, and proportionality to the identified risks; and areas for continued improvement of mitigations.

This Report includes the following components:

- **Bing overview:** An overview of the Bing service and its key features across three categories (core search, generative AI features, and ancillary search features) as well as updates to Bing features during the assessment period.
- **Bing's approach to protecting users and their fundamental rights online:** An overview of (1) the principles Bing follows to provide users with high quality, effective, and safe services, (2) Bing's risk profile and key risk considerations, (3) Bing's approach to mitigating risks on core search, generative AI Features, and ancillary search features, and (4) the methods Bing implements to monitor the effectiveness of its mitigations.
- **Systemic Risk Assessment methodology:** An overview of the process Bing used to conduct this year's Systemic Risk Assessment. (Further detail, including risk scoring definitions, criteria, and rating scales, is included in the Appendix I: Detailed Risk Assessment and Scoring Methodology.)
- **Risk Assessment results:** A description of each risk area, the risk analysis conducted, and the key relevant implemented mitigations.
- **Additional Insights:** A summary of changes to the Bing Risk Assessment team's understanding of Bing's risk profile this year, the enhancements made to each of the priority safety mitigations outlined in last year's report, and the areas Bing will focus on improving for the next reporting period.
- **Conclusion:** A summary of the key considerations impacting Bing's Systemic Risk Assessment results and resulting in an overall low residual risk score.

- **Appendix I - Detailed Assessment Methodology:** Detailed explanation of the methodology used for this year's Systemic Risk Assessment, including the risk definitions, inputs, risk scoring criteria, and rating scales.
- **Appendix II - Catalog of Mitigations by Industry Best Practices:** An overview of Bing's cross-risk mitigation measures categorized across the 35 Digital Trust & Safety Partnership (DTSP) best practices.
- **Appendix III – Abbreviated Terms:** Definitions of the acronyms and initialisms mentioned throughout the report.

How to read this document

At a high level, this document is structured to guide readers to understand Bing and its features; Bing's approach to protecting users and their fundamental rights online; the process and methodology used to evaluate systemic risks in Bing; the results of the evaluation of risks in Bing; and Bing's approach to addressing the residual risks as prioritized by the evaluation. This structure prioritizes explaining the Bing service and its safety protections, then explaining how risks manifest in Bing, and how the safety protections address or need improvement in order to sufficiently address risks to fundamental rights.

Other readers may be seeking a different prioritization of information from this report. Here are suggested options for reading this report:

- **Readers prioritizing understanding Bing's features and approach to protecting users and their fundamental rights online.** Read document as structured.
- **Readers prioritizing understanding Bing's risk assessment methodology and results.** Read document as follows:
    - Systemic Risk Assessment methodology
    - Appendix II: Detailed risk assessment and scoring methodology
    - Risk assessment results
    - Bing's approach to protecting users and their fundamental rights online
    - Bing overview
    - Conclusion
- **Readers prioritizing understanding Bing's improvements from last year.** Read document as follows:
    - Conclusion
    - Additional insights
    - Systemic Risk Assessment methodology
    - Risk assessment results
    - Bing's approach to protecting users and their fundamental rights online
    - Bing overview
- **Readers prioritizing understanding how risks may manifest in Bing.** Read document as follows:
    - Risk assessment results
    - Additional insights
    - Systemic Risk Assessment methodology
    - Bing's approach to protecting users and their fundamental rights online
    - Bing overview
    - Conclusion

Bing overview

As an online search engine, Bing's primary objective is to discover, understand, and organize the Internet's content to offer the most relevant and high quality results available in response to user queries. Bing supplements this core functionality with additional features designed to help users find answers to their questions more efficiently, such as enhanced search features like answers and search suggestions; narrowed search verticals like maps, shopping, travel, or video; and generative AI features like Copilot in Bing and Image Creator from Bing[1].

In line with Microsoft policies and principles around responsible AI, privacy, digital safety, information integrity, and other critical social issues, Bing has developed a holistic digital safety ecosystem that encompasses programmatic safety systems involving ranking improvements, content filtering, abuse detection, threat intelligence monitoring, algorithmic transparency, user controls and reporting, and specialized safety infrastructure for Copilot in Bing and Image Creator from Bing, among other protections, to provide a safe search experience for its users throughout the service. Bing's comprehensive safety framework is further described in Bing's approach to risk mitigation.

User base and market

Bing had approximately 124 million average monthly active users (MAU) in the EU during the period of August 2023 to June 2024. This figure reflects users located within the EU (as identified by IP address) who - at least once per month - query Bing and view the results page on a desktop PC.

Bing has both authenticated and unauthenticated users. With the exception of certain optional features that require authentication, Bing features are available to both authenticated and unauthenticated users.

Most of Bing's services are available to users across ages, subject to parental consent requirements. Authenticated users under the age of consent in their local region require parental consent to use the service. Certain features, including Copilot in Bing, are not available to minor users under the age of consent in the local market (minimum age 13 globally).

Bing's average monthly active user base is not necessarily an accurate proxy for the reach or impact of Bing, as a sizable portion of Bing users are "light users" that query Bing a few days per month. Bing research suggests this can include users who may primarily use other search engines (or social media in lieu of a traditional search engine) but, who once or twice a month enter queries into a Bing interface, such as through the Windows Start menu or Microsoft Edge browser. This is reflected by the meaningfully

[1] While these generative AI features were originally known as "Bing Chat" and "Image Creator", these features have evolved into a new distinct family of AI services under the brand Microsoft "Copilot" during the Reporting Period. Microsoft Copilot acts as an AI-powered research assistant, planner, and creative partner for users across the Microsoft ecosystem and is as a distinct, standalone service with associated endpoints in other Microsoft services to help users complete tasks in a variety of productivity scenarios. Bing Search integrates Copilot functionality to provide users with a modern, natural language-based search interface within the traditional Bing search engine experience. Copilot as it appears in Bing is referred to "Copilot in Bing" herein. Although Copilot is a distinct product, Bing's risk assessment includes information about how risks pertaining to Copilot in Bing are identified and mitigated in the interests of providing a fulsome overview of possible risks that can arise on the Bing platform. This report is directed exclusively at the versions of these features that are available on Bing.

lower amount of daily active users in the EU, which was approximately 18 million for the six-month period ending June 30, 2024. At the same time, Microsoft recognizes the potential for content risks with this smaller population of users and Bing takes steps commensurate with its risk profile and its status as a designated VLOSE to address digital safety for users in the EU and beyond.

Bing offers over 200 market versions and is available in over 100 different languages, including in all EU Member States as of August 2024.

Core search

Feature summary

Core search refers to the "core" Bing search engine experience wherein users can enter queries to search the Bing index for relevant web results. Bing crawls the web to build an index of pages (or URLs) to display as a set of search results relevant to a user-initiated search or action. Complex algorithms generate Bing search results by matching the user's search query with third-party webpages in Bing's index. In some scenarios, Bing may suggest search topics for a user based on their search history or trending topics, such as via search suggestions, and homepage content on the Images and Videos tabs. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to reduce the risk of users encountering harmful or misleading content (absent explicit intent to look for such content). Bing also removes illegal content such as child sexual exploitation and abuse imagery (CSEAI).

Given the vast number of websites on the Internet and the trillions of websites in the Bing search index, the content of these pages may vary greatly and can include images, videos, audio, text, links, downloadable documents, or other materials. Bing does not host the content on the websites that appear in search results, but at most caches third-party pages in order to deliver search results more quickly to users. Bing has no control over the operation, design, or contents of the materials on third-party websites, and third-party websites are not content provided by "recipients of the service" within the meaning of the DSA Article 3 definition. As long as a website makes content available on the Internet and to search engine crawlers, the content will generally be available through Bing and other search engines. Bing does not allow users to post or share their own content on the service; rather, it allows users to find and consume information.

Core search may also provide users with additional features to help provide additional context and information and enhance the search experience. For example, if users want to know which team won a particular sports match, when users search on Bing for a team name, they could be presented with a special Answer including the final score and a recap of the most recent match at the top of the page, as well as an overview with more information about the team, including news, videos, and related searches along with the most relevant search results. To help users find what they are looking for more easily, Bing also lets users narrow their search results with categories like images, videos, news, and shopping. These features also rely on the ranking principles and main parameters of core search, but in some cases take additional considerations into account to provide users with the most relevant results.

How Search works

Returning search results involves complex, near-instantaneous algorithmic calculations. The first step in building Search is figuring out which pages exist on the Internet in order to index them, which is a process commonly called crawling. Pages identified through Bing's crawler are added to the Bing index and algorithms are used to analyze the pages to effectively include them in search results, including determining which sites, news articles, images, or videos will be included in the index and available when users search for specific keywords.

The primary parameters Bing uses to rank pages in search results, in general order of importance, are:

- **Relevance:** Relevance refers to how closely the content on the landing page matches a user's intent behind a search query. Bing presumes the user seeks high quality, authoritative content unless the user clearly indicates an intent to research low quality content.

- **Quality and Credibility:** Determining the quality and credibility (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?
  - **Level of discourse:** Is the page presented in a logical manner and free of clear grammatical errors? Does the page resort to name calling? Does the page advocate violence in any way?
  - **Level of distortion:** How well does the site differentiate fact from opinion?
  - **Origination and transparency of ownership:** Is the site reporting first-hand information or does it summarize or republish content from others? If the site does not publish original content, do they attribute the source?

- **User engagement:** Bing also considers how users interact with search results. To determine user engagement, Bing asks questions like: Did users click through to search results for a given query, and if so, which results? Did users spend time on the search results they clicked through, or did they quickly return to Bing? Did the user adjust or reformulate their query?

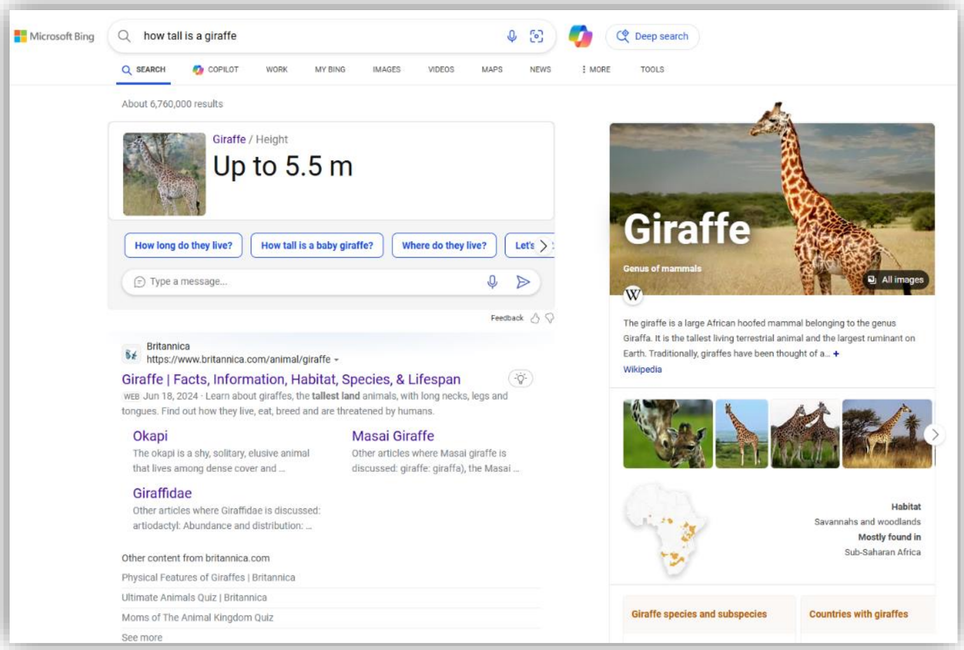
- **Freshness:** Generally, Bing prefers fresh content. A page that consistently provides up-to-date information is considered fresh.

- **Location and Language:** In ranking results, Bing considers the user's location (country and city), the location where the page is hosted, the language of the page, and the location of other visitors to the page.

Answers

"Answers" are enhanced search results provided at the top or side of standard search results that provide richer content in response to a user's query. For example, if a user types "How tall is a giraffe?" Bing will respond with the answer of "up to 5.5m" to enable users to quickly find the information they need, along with a sidebar pane with more information about giraffes. Bing derives answers from high authority search results across the web and links to the original source. Bing may also offer specialized answers, such as in relation to elections or other high interest topics. Some of these answers may be powered by generative AI to provide an even richer experience, linking to key sources.

Figure 1: Core search results screenshot



how tall is a giraffe

About 6,760,000 results

Giraffe / Height

Up to 5.5 m

Britannica
https://www.britannica.com/animal/giraffe

Giraffe | Facts, Information, Habitat, Species, & Lifespan

WEB Jun 18, 2024 · Learn about giraffes, the **tallest land** animals, with long necks, legs and tongues. Find out how they live, eat, breed and are threatened by humans.

Okapi
The okapi is a shy, solitary, elusive animal that lives among dense cover and ...

Masai Giraffe
Other articles where Masai giraffe is discussed: giraffe: giraffa), the Masai ...

Giraffidae
Other articles where Giraffidae is discussed: artiodactyl: Abundance and distribution: ...

Other content from britannica.com
Physical Features of Giraffes | Britannica
Ultimate Animals Quiz | Britannica
Moms of The Animal Kingdom Quiz
See more

Giraffe
Genus of mammals

The giraffe is a large African hoofed mammal belonging to the genus Giraffa. It is the tallest living terrestrial animal and the largest ruminant on Earth. Traditionally, giraffes have been thought of a... +
Wikipedia

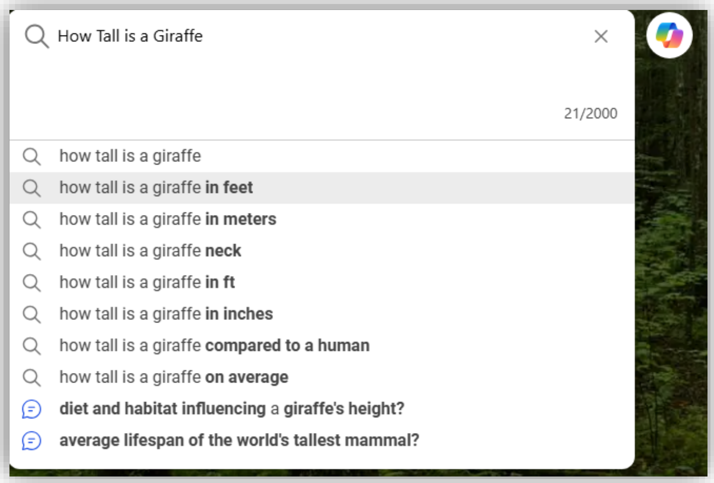
Habitat
Savannahs and woodlands
Mostly found in
Sub-Saharan Africa

Autosuggest and Related Suggestions

Autosuggest and Related Suggestions ("suggestions") are query suggestions provided by Bing to help users use search more conveniently. Suggestions is a feature that shows topics that the users might be interested in next to search results (e.g., "People may also ask").

Suggestions are generated and algorithmically ranked based on the popularity of related searches on Bing and natural language generation technology trained on query sets to help predict a user's intended query, along with other relevance signals such as search history, trends, location, and language.

Figure 2: Search suggestions screenshot



How Tall is a Giraffe

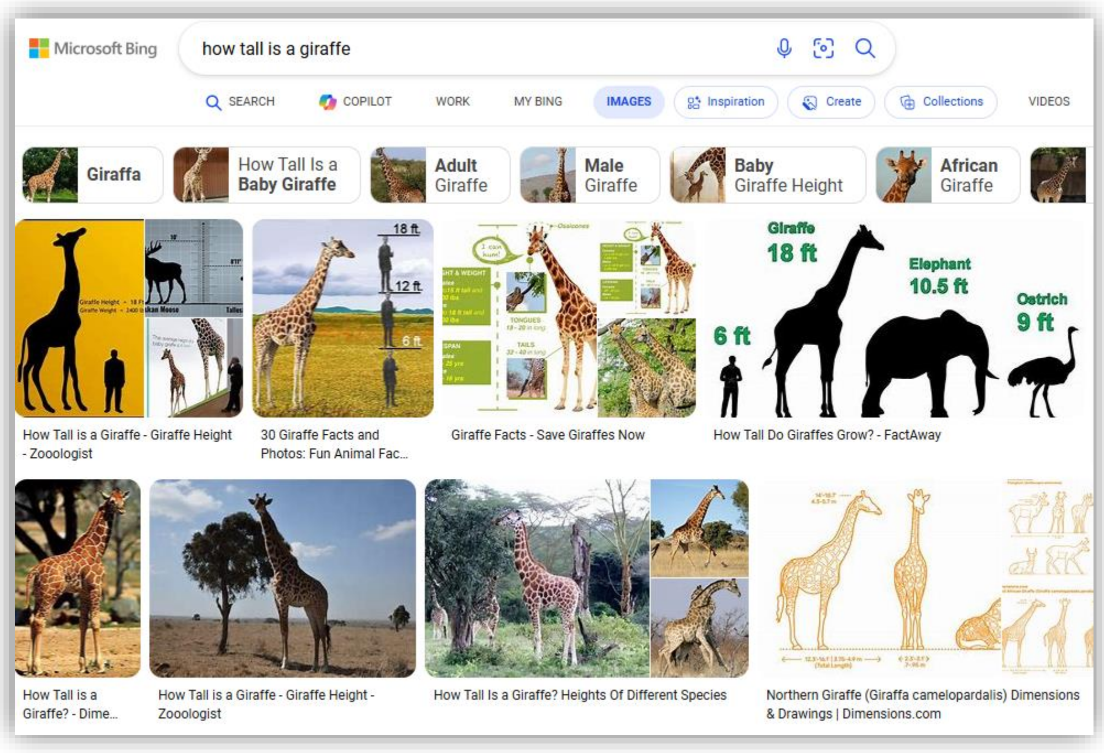
21/2000

- how tall is a giraffe
- how tall is a giraffe **in feet**
- how tall is a giraffe **in meters**
- how tall is a giraffe **neck**
- how tall is a giraffe **in ft**
- how tall is a giraffe **in inches**
- how tall is a giraffe **compared to a human**
- how tall is a giraffe **on average**
- **diet and habitat influencing** a **giraffe's height?**
- **average lifespan of the world's tallest mammal?**

Image and Video

Bing's image and video experiences provide users with image and video results relevant to their search queries. These experiences can appear in the image and video verticals and on the main search results page. Ranking within these experiences generally relies on the same parameters as the main web search results page, e.g., relevance, quality, freshness, authority, and popularity.

Figure 3: Bing Image results screenshot

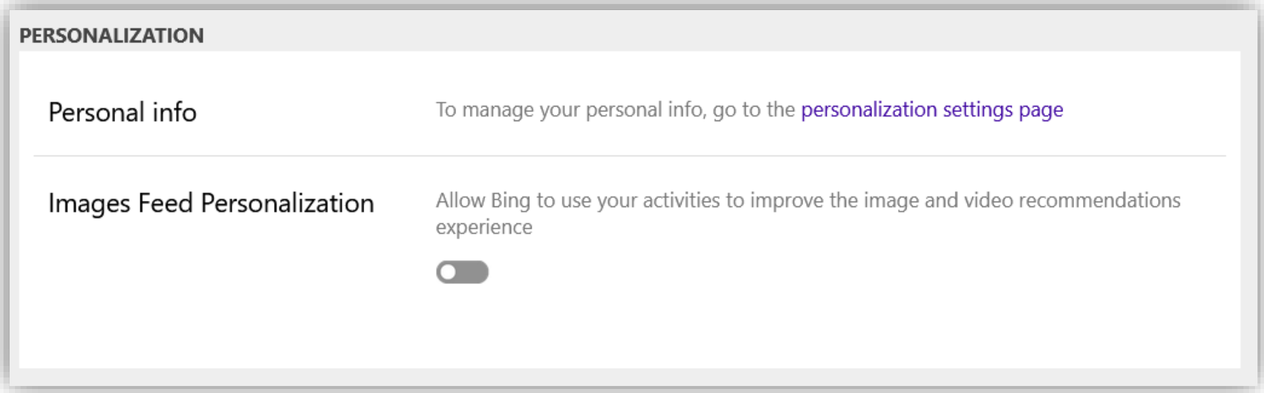


how tall is a giraffe

Microsoft Bing
SEARCH
COPILOT
WORK
MY BING
IMAGES
Inspiration
Create
Collections
VIDEOS
Giraffa
How Tall Is a Baby Giraffe
Adult Giraffe
Male Giraffe
Baby Giraffe Height
African Giraffe
18 ft
12 ft
6 ft
Giraffe 18 ft
Elephant 10.5 ft
Ostrich 9 ft
6 ft
How Tall is a Giraffe - Giraffe Height - Zooologist
30 Giraffe Facts and Photos: Fun Animal Fac...
Giraffe Facts - Save Giraffes Now
How Tall Do Giraffes Grow? - FactAway
How Tall is a Giraffe? - Dime...
How Tall is a Giraffe - Giraffe Height - Zooologist
How Tall Is a Giraffe? Heights Of Different Species
Northern Giraffe (Giraffa camelopardalis) Dimensions & Drawings | Dimensions.com

When a user first lands on the image or video experience homepages, prior to entering a query, the homepage may show recommended content based on prior interactions with the site, such as the user's search history, engagement with image results, saved images, and - where a user allows it - their Edge browsing history. Users can control this experience by deleting data used to influence personalization from their Privacy Dashboard, leveraging the consent for Microsoft Edge browsing activity for personalized advertising and experiences, opting out of personalization in its entirety, or, if they are not logged in, starting a new session. Users can also toggle their image and video feed personalization through Bing's Settings.

Figure 4: Personalization toggle screenshot



PERSONALIZATION

Personal info — To manage your personal info, go to the personalization settings page

Images Feed Personalization — Allow Bing to use your activities to improve the image and video recommendations experience

Generative AI features

Feature summary

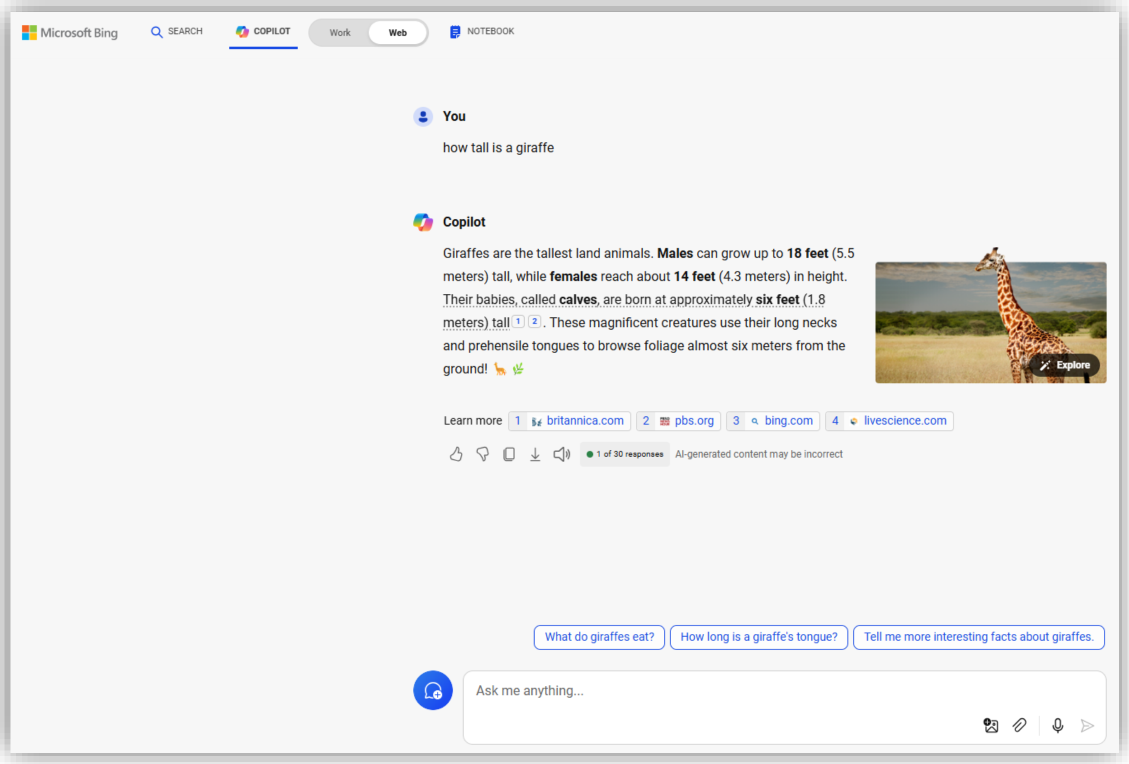
Generative AI features offered via Bing include Copilot in Bing and Image Creator from Bing. These AI experiences are in part powered by Large Language Models (LLMs) developed by Microsoft's technology partner Open AI, including GPT, a cutting-edge LLM, and DALL-E, a deep learning model to generate digital images from natural language descriptions. Microsoft generative AI tools, including Copilot in Bing and Image Creator from Bing, are operated independently of OpenAI, and Microsoft has added additional safety mitigations and modifications to the underlying models, including combining with search algorithms, adding a bespoke metaprompt to guide generative AI responses, and additional safety filters and classifiers.

Responses are presented to users in several different formats, such as traditional links to web content, AI-generated summarizations, images, and chat responses. Summarizations and chat responses that rely on web search results will include references and a "Learn more" section below the responses, with links to search results that were used to ground the response. Users can click these links to learn more about a topic and the information relied upon in the summary or chat response.

With Copilot in Bing, users can request information conversationally by adding context to their prompt (or query) and interacting with the system responses to further specify their search interests. For example, a user might ask follow-up questions, request additional clarifying information, or respond to the system in a conversational way. In the chat experience, users can also select a response from pre-written suggestions, which Bing refers to as chat suggestions. These buttons appear after each response from Copilot in Bing and provide suggested prompts to continue the conversation within the chat experience. For example, if a user is asking about trips to Paris, Copilot in Bing may display a chat suggestion that says: "Tell me about the Louvre Museum." Chat suggestions also appear alongside summarized content on the search results page as an entry point for the chat experience. Bing also allows a user to create stories, poems, song lyrics, and other creative works. When Bing detects user intent to generate creative content (for example, the prompt begins with "write me a …") within the Copilot in Bing experience, the system will generate content responsive to the user's prompt.

In Visual Search in Copilot in Bing, with an image taken by the user's camera, uploaded from the user's device, or linked from the web, users can prompt Copilot in Bing to understand the context, interpret, and answer questions about the image. Users can also upload their files to Copilot in Bing to interpret, convert, process, or calculate information from them.

Figure 5: Copilot in Bing results screenshot



You

how tall is a giraffe

Copilot

Giraffes are the tallest land animals. **Males** can grow up to **18 feet** (5.5 meters) tall, while **females** reach about **14 feet** (4.3 meters) in height. Their babies, called **calves**, are born at approximately **six feet** (1.8 meters) tall [1] [2]. These magnificent creatures use their long necks and prehensile tongues to browse foliage almost six meters from the ground! 🦒 🌿

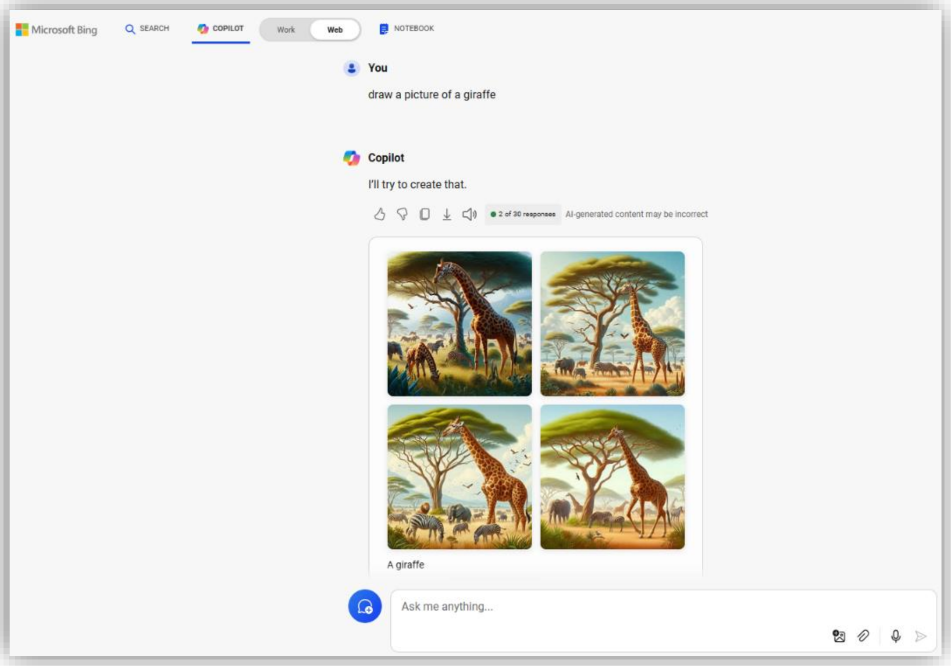
Learn more 1 britannica.com 2 pbs.org 3 bing.com 4 livescience.com

● 1 of 30 responses AI-generated content may be incorrect

Explore
What do giraffes eat?
How long is a giraffe's tongue?
Tell me more interesting facts about giraffes.
Ask me anything...

When Copilot in Bing detects user intent to generate an image (for example, the prompt begins with "draw me a …"), the system will, in most cases, generate an image responsive to the user's prompt using the Image Creator from Bing technology. By typing in a description of an image and other context, such as artistic style or key image details desired, Image Creator from Bing will generate a response, which is then displayed in Copilot in Bing. For example, a user can type "create an image of a giraffe" as a prompt in Copilot in Bing to trigger Image Creator from Bing to generate graphics depicting a giraffe. Image Creator from Bing relies upon an image generation model from OpenAI, with additional Microsoft safety measures implemented. With Image Creator from Bing, users can not only generate images using prompts but can also resize or restyle them by choosing from different restyle options.

Figure 6: Image Creator from Bing results screenshot



You

draw a picture of a giraffe

Copilot

I'll try to create that.

● 2 of 30 responses AI-generated content may be incorrect

A giraffe

Ask me anything...

Copilot in Bing textual responses are generally grounded in web search results. “Grounding” refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material. Copilot in Bing acts like a personal research assistant—users ask the tool a question, and Copilot in Bing combs through relevant search results to provide a summarized answer.

Because Copilot in Bing generally returns links alongside its AI-generated responses, users can see sources from which the information in the response originates and evaluate the credibility of a source, just as they do with traditional search.

In some instances, Microsoft may restrict Copilot in Bing or Image Creator from Bing responses that are inconsistent with Bing’s policies and/or applicable law, and Copilot in Bing may decline to respond or direct users to traditional web search.

Ancillary search features

Feature summary

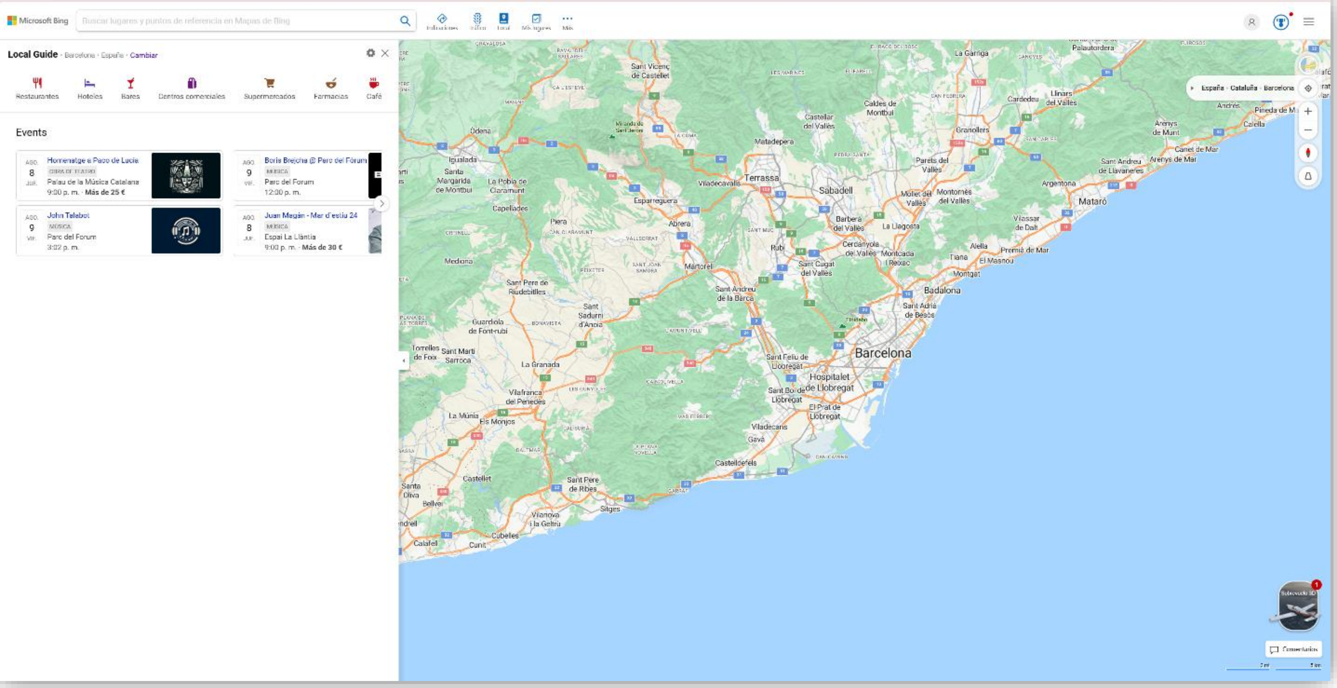
Bing core search links to a number of "verticals" or feature areas to enable enhanced search functionality, including:

- **Maps**
- **Shopping**
- **News**
- **Travel**
- **Real Estate**

Maps

Bing Maps allows users to search for information about businesses, landmarks, locations, and other geographic data in a map interface. In certain jurisdictions, Bing Maps may offer a Local Guide feature where users can view hotels, bars, things to do, restaurants, and events. Businesses can improve the accuracy of Bing Maps results by registering with [Bing Places]() and providing up-to-date information for Bing users on information relevant to that business, such as location, contact information, and hours. Users can also suggest locations to be added to Maps, which then undergo automated and potentially manual review before being made visible to other users. Users can post photos to businesses, which also undergo automated reviews for harmful content.

Figure 7: Bing Maps interface screenshot



AGO. 8 JUE. Homenatge a Paco de Lucía — OBRA DE TEATRO — Palau de la Música Catalana — 9:00 p. m. · Más de 25 €

AGO. 9 VIE. Boris Brejcha @ Parc del Fórum — MÚSICA — Parc del Forum — 12:00 p. m.

AGO. 9 VIE. John Talabot — MÚSICA — Parc del Forum — 3:02 p. m.

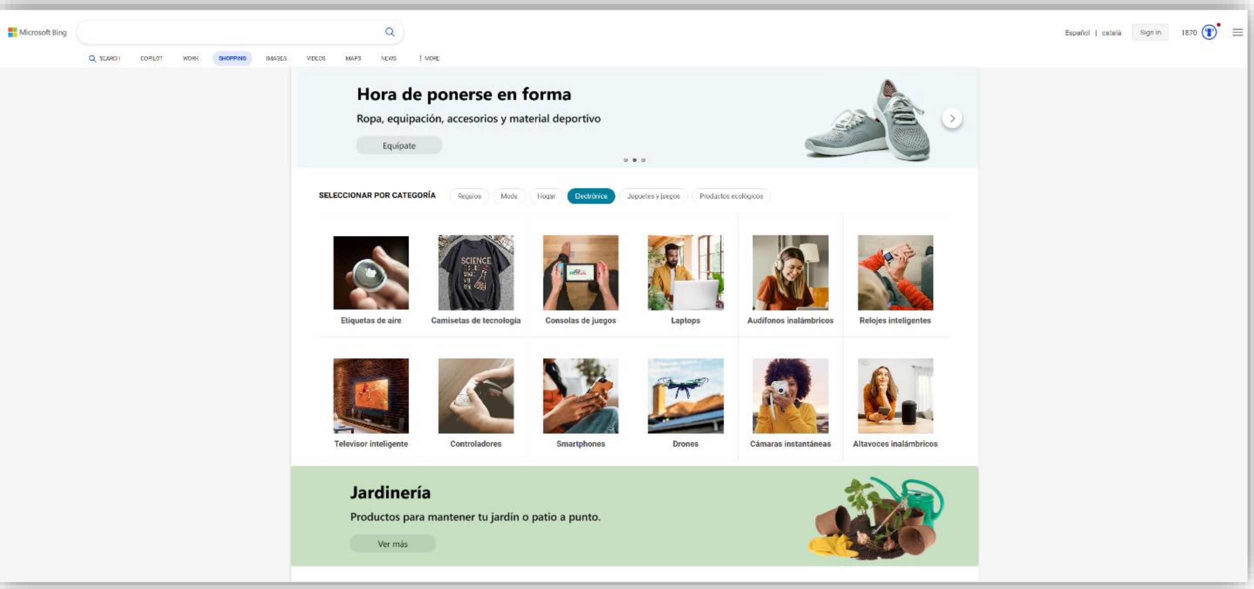
AGO. 8 JUE. Juan Magán - Mar d'estiu 24 — MÚSICA — Espai La Llàntia — 9:00 p. m. · Más de 30 €

Shopping

Bing Shopping is a search feature to help users discover products available through third-party websites by returning relevant shopping results from the web as well as sponsored advertisements relevant to a user's query. As with web results, users in the EU must click through to the third-party websites offering products for sale to complete a purchase. Users do not make purchases directly on or from the Bing service, and Bing does not process payment. In some cases, the shopping experience may use generative AI tools to provide better results for users. In some instances, Microsoft may receive compensation for users who click on algorithmically generated results, but such compensation has no effect on the ranking or relevance of algorithmically generated results shown to the user.

Users who allow Microsoft to collect and use personal data to personalize advertisements may see results in the Shopping vertical tailored to their interests. Users can visit the Microsoft Privacy Dashboard or the Edge shopping toggle to opt-in or out of personalized advertisements and control the data that influences personalization.

Figure 8: Bing Shopping landing page screenshot



Hora de ponerse en forma

Ropa, equipación, accesorios y material deportivo

Jardinería

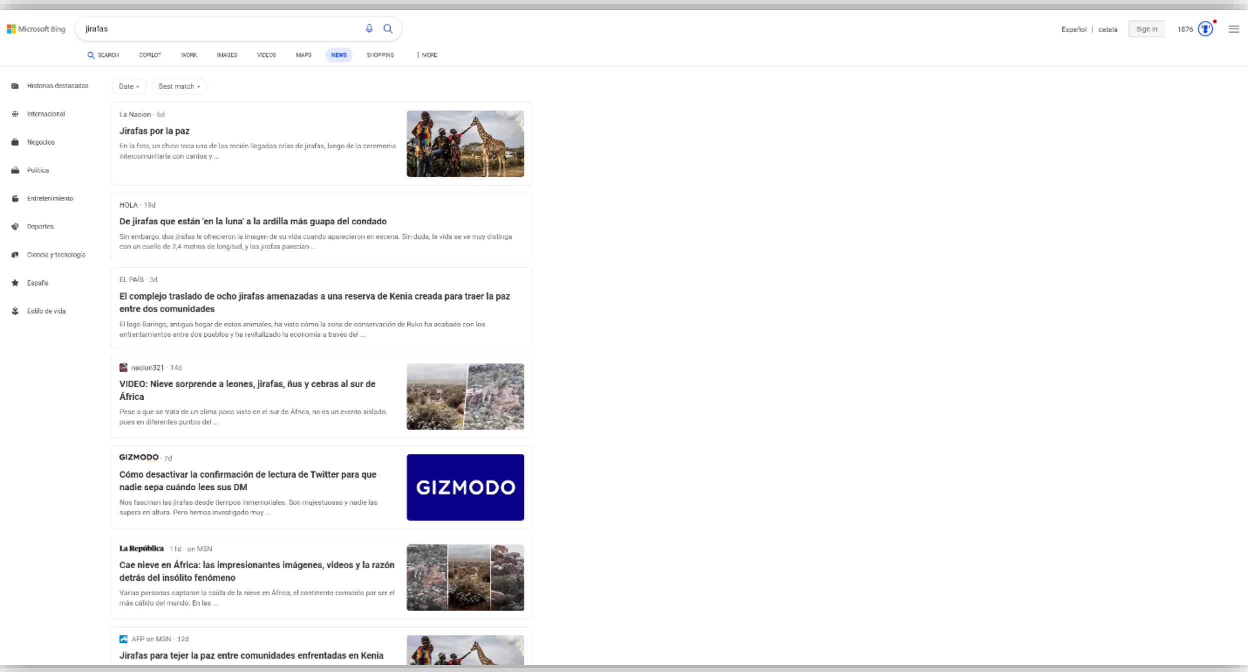
Productos para mantener tu jardín o patio a punto.

News

Bing News is a dedicated news vertical which delivers high authority news content for Bing users, including news search results, news answers, and news carousels.

While Bing News relies on the same main parameters for ranking results as Bing, content that appears in Bing news search results must meet the Bing News Publisher Guidelines, described in the Bing News Publication Hub. News sites interested in being included in the Bing News index may apply through the Publication Hub.

Figure 9: Bing News search vertical



La Nacion · 6d

Jirafas por la paz

En la foto, un chico toca una de las recién llegadas crías de jirafas, luego de la ceremonia intercomunitaria con cantos y ...

HOLA · 19d

De jirafas que están 'en la luna' a la ardilla más guapa del condado

Sin embargo, dos jirafas le ofrecieron la imagen de su vida cuando aparecieron en escena. Sin duda, la vida se ve muy distinga con un cuello de 2,4 metros de longitud, y las jirafas parecían ...

EL PAÍS · 3d

El complejo traslado de ocho jirafas amenazadas a una reserva de Kenia creada para traer la paz entre dos comunidades

El lago Baringo, antiguo hogar de estos animales, ha visto cómo la zona de conservación de Ruko ha acabado con los enfrentamientos entre dos pueblos y ha revitalizado la economía a través del ...

nacion321 · 14d

VIDEO: Nieve sorprende a leones, jirafas, ñus y cebras al sur de África

Pese a que se trata de un clima poco visto en el sur de África, no es un evento aislado, pues en diferentes puntos del ...

GIZMODO · 7d

Cómo desactivar la confirmación de lectura de Twitter para que nadie sepa cuándo lees sus DM

Nos fascinan las jirafas desde tiempos inmemoriales. Son majestuosas y nadie las supera en altura. Pero hemos investigado muy ...

La República · 11d · on MSN

Cae nieve en África: las impresionantes imágenes, videos y la razón detrás del insólito fenómeno

Varias personas captaron la caída de la nieve en África, el continente conocido por ser el más cálido del mundo. En las ...

AFP on MSN · 12d

Jirafas para tejer la paz entre comunidades enfrentadas en Kenia

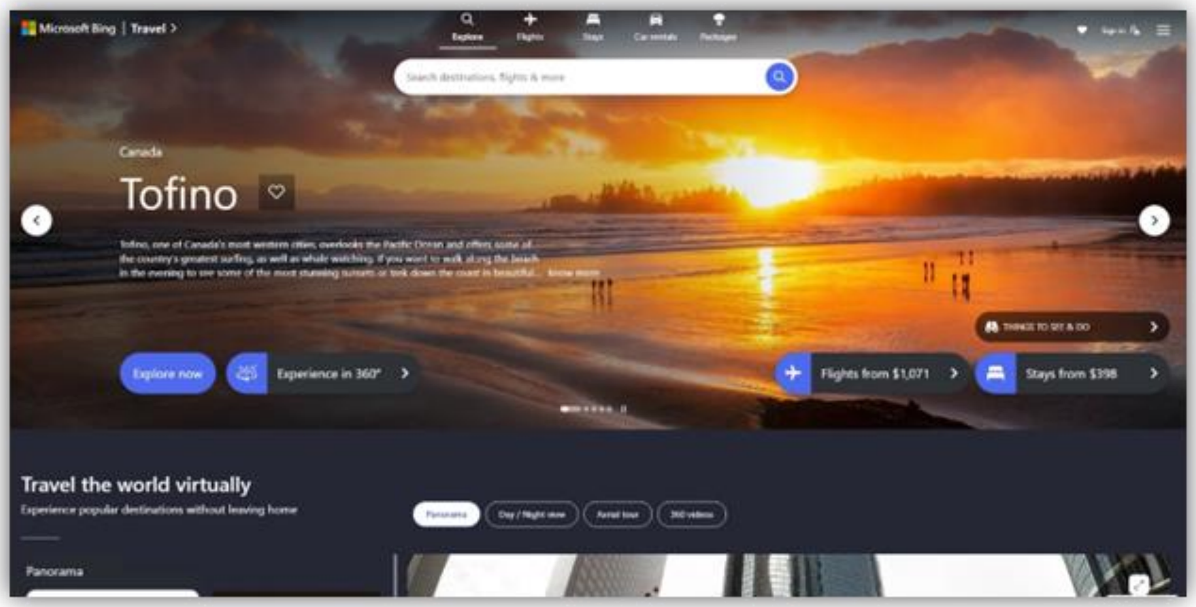
Travel

Bing Travel is a specialized search and trip-planning tool that helps users find information about travel destinations, flights, accommodation, car rentals, and travel packages offered on third party websites. Users can search for flights, hotels, and other travel information. While Bing may display pricing and other key details, users complete all bookings through external booking websites either via links to third party travel websites or through embedded booking tools operated by third party travel partners. Bing does not manage the payment or booking transactions.

Bing Travel may also suggest activities under “What to see & do” for the desired trip and promote other travel-related content.

Bing Travel also includes “Travel Stories” for select locations, featuring third party photo and video content curated for Bing by a third-party content partner.

Figure 10: Bing Travel home page screenshot

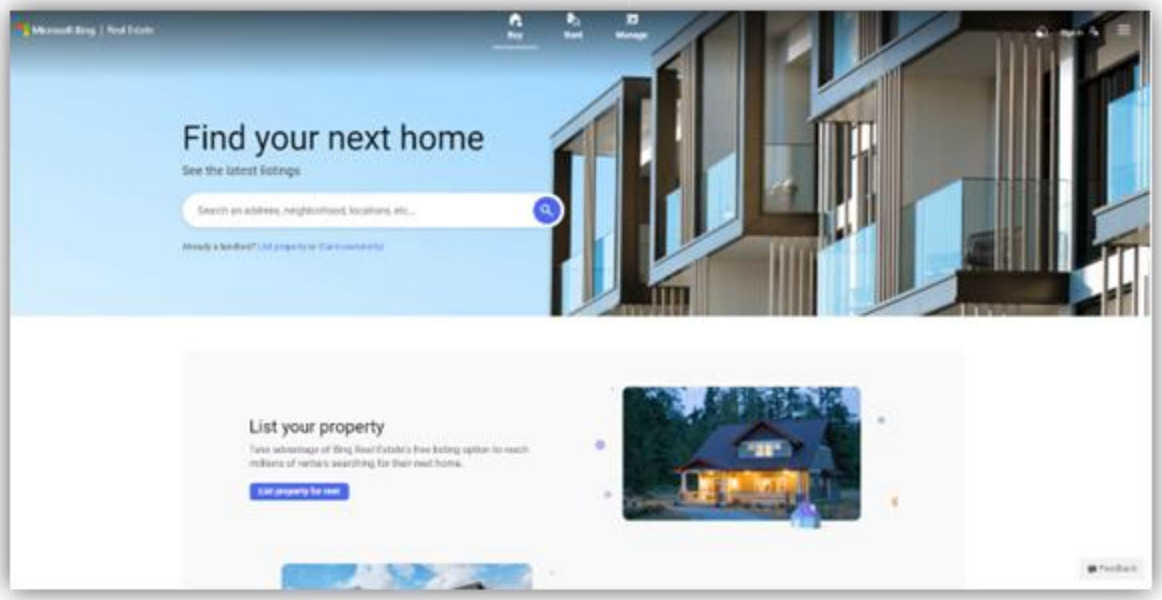


Tofino, one of Canada's most western cities, overlooks the Pacific Ocean and offers some of the country's greatest surfing, as well as whale watching. If you want to walk along the beach in the evening to see some of the most stunning sunsets or trek down the coast in beautiful... Know more

Real Estate

Bing Real Estate offers users the opportunity to search for and view homes for sale or rent in their desired locations and to manage property rentals. Sale and rental results primarily come from partner property sites; however, within the management feature, authenticated users may post properties for rent on Bing real estate and receive and respond to communications from prospective and current renters.

Figure 11: Bing Real Estate search home page screenshot


Microsoft Bing | Real Estate
Buy
Rent
Manage
Find your next home
See the latest listings
Search an address, neighborhood, locations, etc...
Already a landlord? List property or Manage
List your property
List property for rent
Feedback

Take advantage of Bing Real Estate's free listing option to reach millions of renters searching for their next home.

Advertisements appearing on Bing

Bing, as a free service, is monetized almost entirely through advertising. When a user conducts a search on Bing, the user is shown both “organic” (i.e., unpaid) results and, when relevant, advertisements (i.e., paid advertising). Advertisements across Bing are clearly labeled as “Ads” and are displayed to users primarily based on their search queries. Advertisements that appear on Bing are provided by Microsoft Advertising, which is a business-to-business service separate from Bing that delivers advertising to a range of Microsoft services as well as third-parties. As a separate service, advertisements on Bing are governed by the Microsoft Advertising Agreement, including the Microsoft Advertising Network Policies.

New Bing features

Bing endeavors to provide innovative experiences for its users and regularly evaluates the performance and usefulness of its features. As a result, features may be modified, tested, removed, improved, or adjusted to improve or iterate on the Bing experience. Throughout the year, Bing tracks updates to features, including rollbacks, changes, and launches through its launch readiness review. New features and in-scope feature updates to Bing must go through this launch readiness review.

This review process requires product teams that seek to make meaningful changes to these features to conduct a number of different tests and evaluations and complete corresponding documentation of the testing and evaluation. Among other things, the product team must have conducted metrics testing of the change and must provide data demonstrating whether the proposed change would result in improvements, regression, or no change against these metrics, and explain any additional mitigations implemented to address identified issues. The review also confirms that additional reviews (such as for digital safety, security, privacy, and accessibility) have been completed prior to launch. Any significant new generative AI feature also goes through a responsible AI launch readiness process to be reviewed and approved by a group of key AI stakeholders across the company.

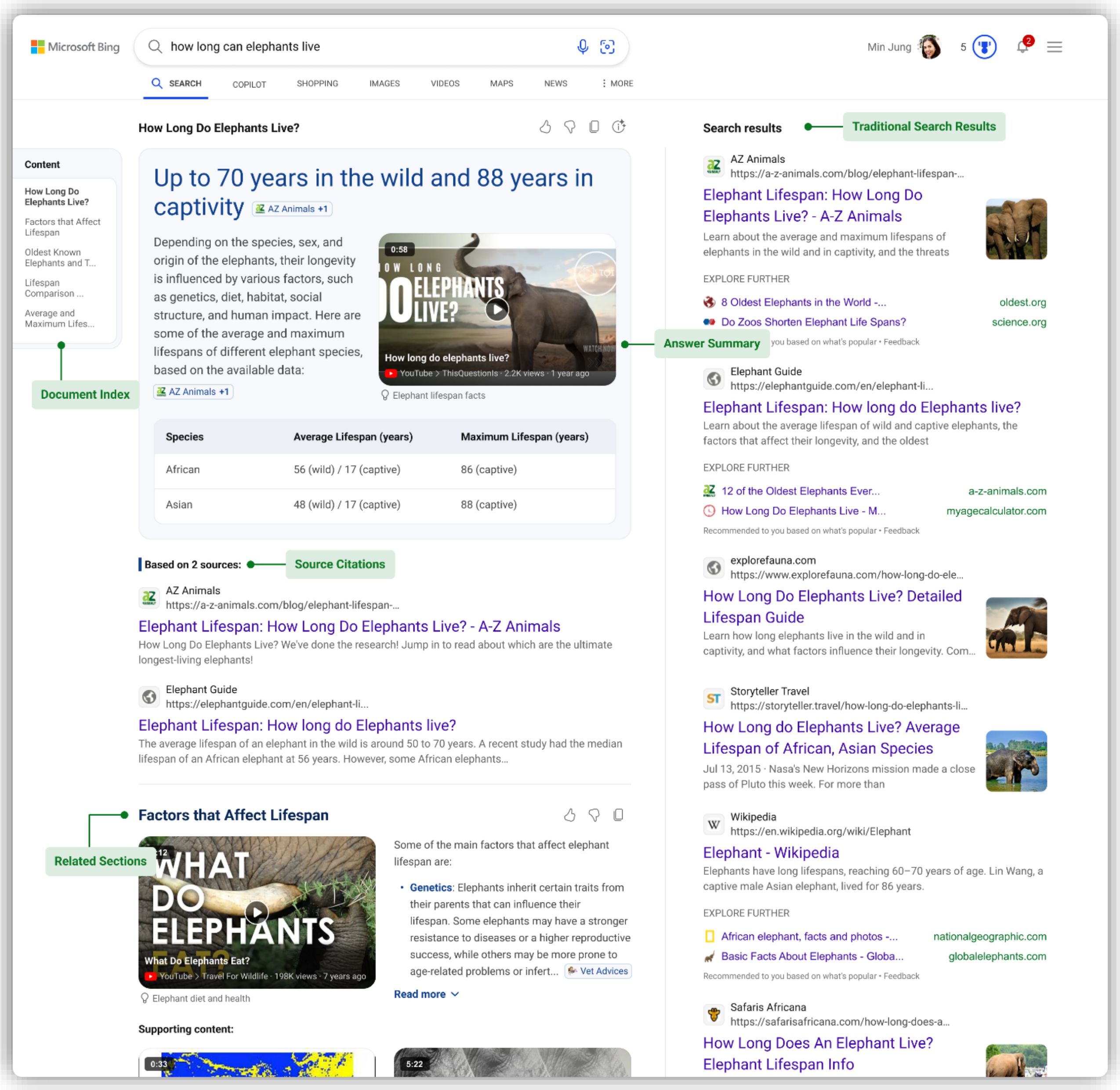
As part of the pre-launch review processes, a core Bing team also evaluates whether features are likely to have a critical impact on systemic risks identified in the DSA, in which case the Bing Risk Assessment team evaluates to ensure adequate mitigations are in place prior to the launch of the updated feature. While no launches during the Reporting Period met the criteria for an out-of-cycle update to the Systemic Risk Assessment, in the interests of transparency several feature and system updates that have occurred since the last Systemic Risk Assessment report are discussed below.

Deep Search

“Deep Search” or “Generative Search” is a feature released during the assessment period that allows users to choose to allow Bing to spend more time looking for, formulating, and returning search results. By asking the user for permission (e.g., click a button) to allow more time (typically up to 20 seconds) to consider their query, Bing can deliver even more relevant results. Deep Search combines the foundation of Bing’s search results with the power of large and small language models (LLMs and SLMs). It aims to understand the search query, review millions of sources of information, dynamically match content, and generate search results in a user-friendly layout that aims to fulfill the intent of the user’s query more effectively. The regular search results continue to be prominently displayed on the page as always.

Deep Search is available through multiple Bing entry points. As part of the user experience, Deep Search can suggest the intent of the query to the user (and present alternative choices if there are multiple potential intents), report the progress and depth of the search, and provide information overviews as part of a more dynamic whole page experience containing relevant results.

Figure 12: Deep Search results screenshot



How Long Do Elephants Live?

Content

How Long Do Elephants Live?

Factors that Affect Lifespan

Oldest Known Elephants and T...

Lifespan Comparison ...

Average and Maximum Lifes...

Up to 70 years in the wild and 88 years in captivity AZ Animals +1

Depending on the species, sex, and origin of the elephants, their longevity is influenced by various factors, such as genetics, diet, habitat, social structure, and human impact. Here are some of the average and maximum lifespans of different elephant species, based on the available data: AZ Animals +1

How long do elephants live?
YouTube › ThisQuestionIs · 2.2K views · 1 year ago

Elephant lifespan facts

Species	Average Lifespan (years)	Maximum Lifespan (years)
African	56 (wild) / 17 (captive)	86 (captive)
Asian	48 (wild) / 17 (captive)	88 (captive)

Based on 2 sources:

AZ Animals
https://a-z-animals.com/blog/elephant-lifespan-...
Elephant Lifespan: How Long Do Elephants Live? - A-Z Animals
How Long Do Elephants Live? We've done the research! Jump in to read about which are the ultimate longest-living elephants!

Elephant Guide
https://elephantguide.com/en/elephant-li...
Elephant Lifespan: How long do Elephants live?
The average lifespan of an elephant in the wild is around 50 to 70 years. A recent study had the median lifespan of an African elephant at 56 years. However, some African elephants...

Factors that Affect Lifespan

What Do Elephants Eat?
YouTube › Travel For Wildlife · 198K views · 7 years ago

Elephant diet and health

Some of the main factors that affect elephant lifespan are:

- **Genetics**: Elephants inherit certain traits from their parents that can influence their lifespan. Some elephants may have a stronger resistance to diseases or a higher reproductive success, while others may be more prone to age-related problems or infert... Vet Advices

Read more

Supporting content:

Search results

AZ Animals
https://a-z-animals.com/blog/elephant-lifespan-...
Elephant Lifespan: How Long Do Elephants Live? - A-Z Animals
Learn about the average and maximum lifespans of elephants in the wild and in captivity, and the threats

EXPLORE FURTHER

8 Oldest Elephants in the World -...	oldest.org
Do Zoos Shorten Elephant Life Spans?	science.org

you based on what's popular • Feedback

Elephant Guide
https://elephantguide.com/en/elephant-li...
Elephant Lifespan: How long do Elephants live?
Learn about the average lifespan of wild and captive elephants, the factors that affect their longevity, and the oldest

EXPLORE FURTHER

12 of the Oldest Elephants Ever...	a-z-animals.com
How Long Do Elephants Live - M...	myagecalculator.com

Recommended to you based on what's popular • Feedback

explorefauna.com
https://www.explorefauna.com/how-long-do-ele...
How Long Do Elephants Live? Detailed Lifespan Guide
Learn how long elephants live in the wild and in captivity, and what factors influence their longevity. Com...

Storyteller Travel
https://storyteller.travel/how-long-do-elephants-li...
How Long do Elephants Live? Average Lifespan of African, Asian Species
Jul 13, 2015 · Nasa's New Horizons mission made a close pass of Pluto this week. For more than

Wikipedia
https://en.wikipedia.org/wiki/Elephant
Elephant - Wikipedia
Elephants have long lifespans, reaching 60–70 years of age. Lin Wang, a captive male Asian elephant, lived for 86 years.

EXPLORE FURTHER

African elephant, facts and photos -...	nationalgeographic.com
Basic Facts About Elephants - Globa...	globalelephants.com

Recommended to you based on what's popular • Feedback

Safaris Africana
https://safarisafricana.com/how-long-does-a...
How Long Does An Elephant Live? Elephant Lifespan Info

Updates to Image Creator from Bing

Image Creator from Bing integrated DALL-E 3, which is the latest iteration of the image generation AI model developed by Microsoft's technology services provider OpenAI. This shift from DALL-E 2 to DALL-E 3 brings to Image Creator from Bing a variety of enhancements, including enhancements in image quality, more photorealistic images, and greater versatility with creative prompts, allowing for broader artistic styling capabilities

Rebranding of Bing Chat and Bing Image Creator

The features referred to in the prior year Systemic Risk Assessment Report as "Bing Chat," and "Bing Image Creator" were rebranded this year as "Copilot in Bing" and "Image Creator from Bing," respectively. There were no material changes made to the services offered via Bing associated with this name change.

Maps - Commute mobile destinations

In this assessment period, Maps expanded the ability for users to identify destinations within Maps for the purpose of creating a personalized commuting route. Maps' Commute feature previously allowed users to identify their home and work to provide a personalized route. With the addition of the destinations feature, users can now add other destinations to their personalized routes, for example a school or gym, to get directions quickly between their home, work, or other top destinations. Within the Commute feature, users can receive personalized traffic news. The user's location is sent without an identifier and used to fetch relevant traffic news. The location is not stored on a server. Users can also choose to share their Estimated Time of Arrival (ETA) using various 1st and 3rd party apps on their mobile device. Personal data is not shared with these apps unless the user explicitly chooses to share it.

Bing’s approach to protecting users and their fundamental rights online

Bing’s values and commitments

Given that most research today is conducted online, search engines play a vital role in society by promoting the fundamental rights to freedom of expression and information and supporting media pluralism. It is integral to Bing’s design principles to ensure that Bing does not unduly restrict users’ ability to receive and impart information. At the same time, Bing recognizes that other fundamental rights and social interests, such as privacy, safety, upholding democracy, public health, and national security, are also vital to Bing’s users and to a healthy society, and Bing must balance these interests and rights to maximize benefit to users while minimizing possible harms.

In order to provide its users with a high quality, effective, and safe search service that appropriately balances tradeoffs between fundamental rights, Bing follows the below “Trustworthy Search Principles” to guide the product design, experience, algorithms, and mitigation measures that Bing adopts to ensure users’ expectations are met while addressing potential risks or harms arising from use of the service.

Bing aims to provide credible and authoritative results relevant to user queries.

- Bing works to provide the highest quality, authoritative content relevant to users’ queries.
- Bing’s goal is to provide fair, balanced, and comprehensive content. When there are multiple credible perspectives, Bing tries to display them in informative ways.
- When there is no authoritative source, Bing’s goal is to avoid promoting bias or potentially misleading information.
- Bing respects user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less credible, while (as described in more detail below) working to ensure that users are not misled by such search results.

Bing promotes free and open access to information within the bounds of the law and with respect for local law and other fundamental rights, such as privacy and public safety.

- Bing provides open access to as much of the web as possible, but in limited cases it may undertake certain interventions (such as removal of a website or downranking) for instances where the content violates local law or Microsoft’s policies.
- When limiting access to content, Bing strives to ensure its actions are narrowly tailored, so it does not unduly restrict important interests, such as the freedom of expression, open

access to information, and media pluralism, and provides transparency regarding its actions.

Bing takes steps to protect users from harmful and unexpected offensive content.

- Bing recognizes that there are many reasons why someone may want to research or review harmful or controversial content, but also recognizes the importance of ensuring users are not inadvertently misled by or unintentionally shown such content.
- For certain types of content where Bing identifies results that may include harmful or misleading information, Bing may provide supplemental information, such as warnings and public service announcements (PSAs), to inform users about potential risks.
- Bing gives users control over the type of content they encounter in Bing through features such as SafeSearch and Family Safety.
- Absent a clear intent to access specific content, Bing assumes an intent to find high authority results.

Bing is transparent about its principles and practices, as well as its decisions and actions.

- Bing provides users with information about its principles regarding ranking and relevance, and moderation policies.
- When Bing limits access to content, where relevant Bing provides notice to users that content was removed.
- Bing publishes regular transparency reports providing information about the complaints it receives, and the actions taken.

AI systems in Bing are developed and evaluated in accordance with Microsoft's Responsible AI Standard:

- **Fairness:** How might an AI system allocate opportunities, resources, or information in ways that are fair to the humans who use it?
- **Reliability and Safety:** How might the system function well for people across different use conditions and contexts, including ones it was not originally intended for?
- **Privacy and Security:** How might the system be designed to support privacy and security?
- **Inclusiveness:** How might the system be designed to be inclusive of people of any ability?
- **Transparency:** How might people misunderstand, misuse, or incorrectly estimate the capabilities of the system?
- **Accountability:** How can Microsoft create oversight so that humans are accountable and in control?

In addition, Bing is guided by Microsoft's broader corporate mission and commitments to:

- Expand Opportunity
- Earn Trust
- Protect Fundamental Rights
- Advance Sustainability

More details on these commitments are available here.

Impact to risk profile and key risk considerations

Bing has traditionally evaluated its key risk areas based on aspects of the service that are higher vectors for systemic harms than others. Bing's core functionality is to answer users' search questions, primarily by connecting them with high quality third-party web content relevant to their queries. More recently this has expanded to include generative AI features that allow users to find information more efficiently using natural language conversations; however, as with traditional web search (the results of which Copilot in Bing responses are grounded), the primary functionality is to obtain information.

When using a search engine, users can have many valid reasons for wanting to seek out (legal) content that could be problematic or harmful if encountered in other contexts. As one of the few services that can connect individuals with as-yet-undiscovered content on the World Wide Web, Bing plays an important role in upholding the fundamental rights of free expression and access to information. In this context, one of the highest areas of risk is in ensuring that Bing's algorithmic systems can connect users with the content they are seeking, while ensuring that its content policies and practices regarding this third-party content are sufficiently robust such that users are not inadvertently misled or harmed by search results.

Bing also collects data from its users, including search terms and related data like location and language preferences that help provide more useful search results, and recognizes its obligation to comply with applicable data protection laws and Microsoft policies to ensure appropriate stewardship of that data and minimize risks of harms related to misuse of personal data. However, given that Bing generally does not host user content, allow for messaging between users, or allow users to publish content on the platform, data-related practices in Bing tend to be a lower vector for systemic harms than may be present on other platforms. Because Bing generally does not allow users to post or share their own content on the platform, unlike in social media, there is limited (if any) risk related to user behavior toward other users on the platform, and limited (if any) risk on the platform that user-generated content (or improper enforcement of rules related to user generated content) give rise to significant systemic harms.

Bing does not generally prohibit users from entering search terms that indicate an intent to find harmful, offensive, or potentially illegal content as users can have valid research reasons for seeking out various types of content in search, even content that could be harmful or even illegal in other contexts. This means that while Bing does have established terms and conditions that apply to its users, and does enforce those terms where needed – such as in the case of a user attempting to use illegal CSEAI as a search prompt in visual search, or users who violate terms or attempt to jailbreak generative AI features – user behavior on the platform is not a significant vector for harms on search, and enforcement of Bing's terms and conditions plays a relatively minor role in mitigating systemic harms in Bing.

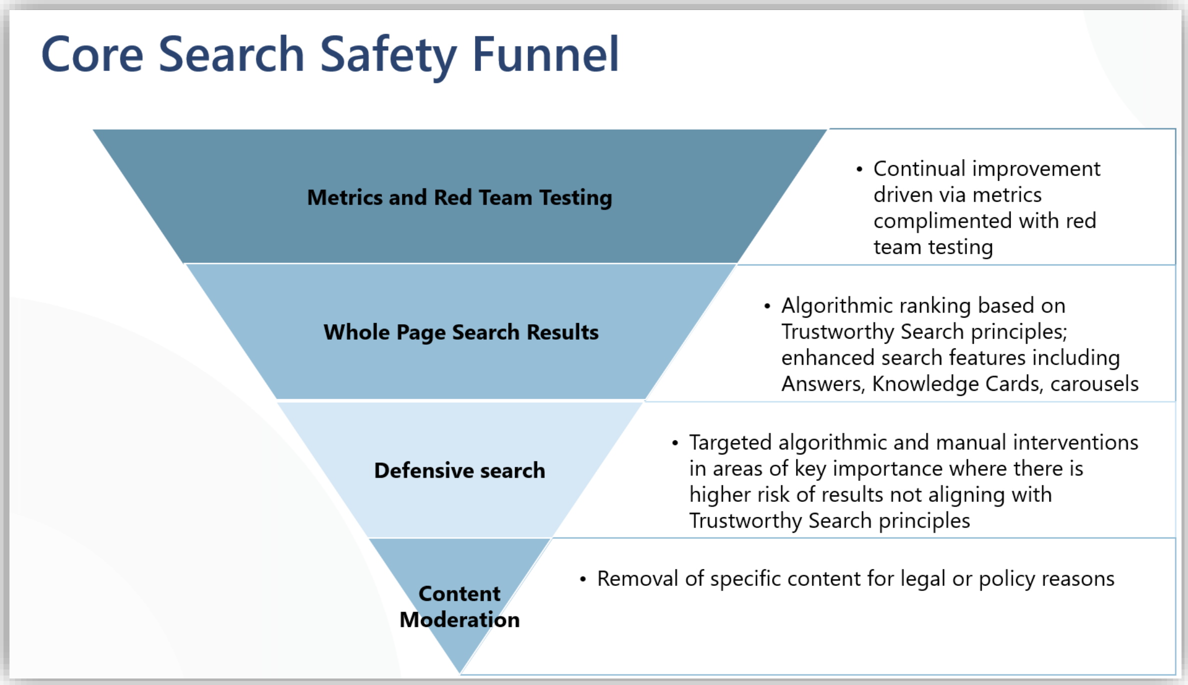
The generative AI features Copilot in Bing and Image Creator from Bing were designed to provide new ways for users to find the information they want: using the power of natural language to make it easier for users to find answers to their research questions, and answer more sophisticated questions, with additional ability to use the tools to inspire new creativity through generating stories, songs, images, or similar outputs. Many of the risks are similar to traditional core Search: ensuring the algorithms and AI systems underlying the service are designed to return the content users are seeking, enhanced by content moderation policies and systems designed to ensure users are not harmed by content they seek in search results. Users still cannot post or share content on these services, but Microsoft recognizes that there is increased risk that users could take content generated using these tools and share it separately on third-party platforms in harmful ways. As a result, Microsoft has implemented specialized AI mitigations (such as metaprompts, filters, and classifiers), enhanced incident response and monitoring, and additional terms and Codes of Conduct (and related enforcement processes) with its generative AI tools to ensure they are used in appropriate ways, and that enforcement actions are fair.

As an ad-supported service, Bing recognizes that there are risks in ensuring that the advertisements appearing on the platform meets Bing's standards for content delivery, in terms of its content policies and enforcement, its ad ranking practices, and data use as well as the standards of Microsoft Advertising.

Service-wide mitigations

In line with these principles, Bing has a multi-layered approach to digital safety that encompasses programmatic safety systems including algorithmic ranking improvement, content removal, incident response, and user feedback and reports, among additional protections to address new possible harms related to AI products to provide a safe experience for its users throughout the service. This approach is described in greater detail in the following feature-specific sections, with some key service-wide mitigations described here to frame Bing's safety strategy. A visualization of how key Bing safety mitigations work together is depicted below.

Figure 13: Core search Safety Funnel



Core Search Safety Funnel

Metrics and Red Team Testing
- Continual improvement driven via metrics complimented with red team testing

Whole Page Search Results
- Algorithmic ranking based on Trustworthy Search principles; enhanced search features including Answers, Knowledge Cards, carousels

Defensive search
- Targeted algorithmic and manual interventions in areas of key importance where there is higher risk of results not aligning with Trustworthy Search principles

Content Moderation
- Removal of specific content for legal or policy reasons

Applicable terms and conditions and their enforcement

Bing's content moderation activities are largely focused on the third-party website content that is linked to from search results[2]; Bing does not control the operation or design of the indexed websites and has no ability to control what those websites publish via terms and conditions. Bing's principles regarding third-party web content are described in How Bing Delivers Search Results and the Bing Webmaster Guidelines. The Microsoft Services Agreement (MSA) serves as Bing's general terms and conditions governing user behavior, with supplemental terms for generative AI features. Unlike in services where user-generated content is core to the service, moderating user content on Bing is not considered key to Bing's safety program because users generally cannot post or share their own content on the service and content from users is not shared through algorithmically-driven recommendation "feeds" to the public or others on the service. Bing generally does not take action against users for their queries (and users may have valid reasons to seek out content in search that could be problematic in other contexts). Bing enforces its terms and conditions where necessary, such as taking action against user accounts where authenticated users attempt to use known CSEAI as a search prompt in visual search.

In addition to the MSA, users of Copilot in Bing and Image Creator are also subject to the Copilot AI Experiences Terms and Image Creator from Microsoft Designer Terms. Users are subject to suspension or other restrictions for the violation of terms and conditions governing user behavior on the service. Microsoft Advertising, which powers ads on Bing, has clear and regularly enforced content policies and practices that prevent advertisements from negatively impacting human dignity. The Microsoft Advertising Policies set out the requirements for ad content, including criteria upon which ad content will be removed. Microsoft requires advertisers to comply with its policies. Microsoft Advertising also has a set of Relevance and Quality Policies to manage the relevance and quality of the advertisements that it serves through its advertising network.

Digital literacy and user control

Bing offers special features and functionalities intended to help foster digital literacy and provides users with options to control their experiences to help them engage with information on the web safely. Bing search results pages may include special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or build special information "hubs" related to important issues containing news and data from high authority sources. Bing also promotes information integrity by supporting trustworthiness signals, such as NewsGuard ratings that allow users to better evaluate their sources of information, and the free and open ClaimReview protocol to embed fact-check tags into articles that appear in search indexes. Copilot in Bing includes in-service disclosures to users that they are engaging with an AI system with reminders that AI can make mistakes. In addition, product FAQs, help pages, and

[2] Note that Bing uses the term "content moderation" broadly in this report to reflect instances where content available on Bing may be removed or otherwise actioned pursuant to its policies or legal obligations, even where such content is not provided by "recipients of the service" within the meaning of the Digital Services Act. Although third-party websites in the Bing index do not constitute content provided by recipients of the Bing service (see Recital 77) and thus cannot be subject to "content moderation" within the meaning of the Article3(t) of the Digital Services Act, the term is used in this report for the purposes of comprehensively discussing risks and broader moderation efforts across the platform.

other public facing information sources help educate users on the nature of AI-driven search experiences and the uses, safeguards, and limitations of this emerging technology, regularly reminding users of the potential for mistakes, need to double check important information, and risks of over-reliance. Bing also provides users with options to adjust their search settings to customize their experience. For example, users can change their SafeSearch setting to prevent the display of adult content in search results; adjust controls that allow them to exercise their data subject rights to view, access, export, and delete personal data held by Microsoft and set location/language controls. Parents can also use Microsoft's Family Safety feature to control the experience of minor's accounts.

Metrics and effectiveness monitoring

Bing monitors and evaluates the effectiveness of Bing's safety systems and many of its mitigations via ongoing metrics monitoring. Bing measures "Defensive Defect Rates" (DDR) to evaluate the performance of and identify gaps in Bing's safety systems to drive improvements. Additionally, Bing performs red team testing pre-launch and post-launch as appropriate to identify issues.

External engagement and consultations

Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as terrorist and violent extremist content, information integrity and misinformation, CSEAI, responsible AI and AI-specific risks, hate speech, minors and technology, and copyright and trademark infringement, to ensure that Bing internal policies, practices, and standards are addressing key concerns of third party stakeholders. These engagements can inform the processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Systems for selecting and presenting advertisements

Ads on Bing are primarily contextually relevant to the search query provided – for example, a search for "flights to Paris" will return advertisements for airlines. Contextual ads are not targeted based on browsing history or interaction with other websites. Advertising customers are governed by Terms and Conditions and Content Policies, which are regularly reviewed and updated to ensure they address key areas of concern and are enforced consistently.

Data-related practices

Bing collects some personal data from users to provide and improve the search services, such as user queries, language preferences, and location. Bing also indexes content from the world wide web, which on occasion contains personal data. Bing has a robust privacy program to ensure it treats personal data in accordance with applicable policies and laws, including impact assessments, transparency, controls, moderation, and data security.

Bing’s approach to risk on core search

Feature-specific risks and mitigations

The Systemic Risk Assessment process considered how the risk profile may differ and how systemic risks may manifest across core search uniquely as well as what feature-specific mitigations are in place to address these risks.

Core search risk profile

As mentioned above, search engines play a vital role in society considering the heavy reliance on search engines to access information that is critical for day-to-day life. In consideration of this role, search engines must carefully balance the fundamental right of freedom of information against digital safety. In some cases, a higher risk tolerance related to safety is warranted to ensure critical access to information.

Users approach search engines with questions and will typically see content that they have requested, rather than content that is shared or proactively recommended to them. Content on a search engine is not generally shared from user to user within the search ecosystem, which reduces the ability for content to “go viral” on the Bing service. Therefore, while referenced harms may be present within core search, the impact is unlikely to become systemic due to the limits on user-to-user sharing and recommendations.

Some ways that systemic risks can manifest on a search engine absent sufficient mitigations include:

- Search results may link users of any age to third-party content that is illegal, unsafe, misleading, fraudulent, private, or otherwise harmful.
- Ranking of search results, search suggestions, or result summaries may perpetuate bias, propagate false or low authority information, contribute to echo chambers, limit pluralism, or normalize harmful content.
- Ranking or demotion of search results may unnecessarily restrict user access to information or limit pluralism.
- Bad actors may exploit Bing to increase potential user exposure to harmful content by manipulating algorithms to increase the prominence of lower quality or authority content.
- Bad actors may breach search engine services to access or expose user queries or other data retained by the search engine.

User Counts

Core search represents Bing’s central function as a search engine and is the source of the vast majority of Bing’s usership with the highest usage of any Bing feature or service.

Overall approach to mitigating risks on core search

Microsoft respects freedom of expression and the right to access information. At the same time, in accordance with Microsoft policies and principles around responsible AI, privacy, digital safety, information integrity, and other critical issues, Bing has developed a safety system including content filtering, operational monitoring, and abuse detection to provide a safer search experience.

Search algorithms and recommender systems are a fundamental part of Bing’s risk mitigation approach. Complex algorithms generate Bing search results by matching the user’s search query with third-party webpages in Bing’s index. The majority of content displayed by Bing is in response to an explicit user query, as opposed to content that is recommendations based on implied user behavior across the platform. In some scenarios (with appropriate controls and mitigations), Bing may suggest search topics for a user based on their search history or trending topics, such as via search suggestions, and homepage content on the Images and Videos tabs. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content and help ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and uses authority as a part of the Quality and Credibility (QC) score, which is one of Bing's main ranking parameters.

Bing continuously measures metrics to evaluate the effectiveness of its safety systems, identify gaps, and inform Bing's mitigation strategy. Although the Bing search algorithms are designed, and continually refined, to prioritize high authority content in the search results, in some cases, additional mitigations are required to address the online safety risks. As such, Bing implements a multi-tiered safety strategy to complement the algorithmic prioritization of high authority content, including the following mitigations.

Detailed descriptions of core search mitigations

Understanding of authority on the web is the foundation of Bing's approach to risk mitigation on core search. Bing collects signals from external entities like NewsGuard, from fact checkers via partnerships, or using open protocol such as ClaimReview tag. These signals are monitored to directly identify some low authority sites which Bing then uses to identify other low authority sites. Similarly, in addition to manual intelligence gathering, Bing monitors known low authority sites to identify new topics related to other relevant threats or emerging risks. Sites identified as low authority will be ranked with less prominence for any query.

Additional ranking interventions

In some cases, Bing may discover that certain search topics are returning results not in line with Bing's principles, such as where bad actors are intentionally manipulating search results to surface low authority content, exploiting a data void, or otherwise violating Bing's Webmaster Guidelines in an effort to improperly manipulate search results. Prohibited practices are described further at Webmaster Guidelines-Bing Webmaster Tools. When this occurs, Bing may apply additional algorithmic interventions to prevent users from being unexpectedly exposed to harmful or misleading content that they did not demonstrate an intent to search for. Bing prioritizes interventions on known misinformation trends, topics where there are elevated risks of active manipulations, and areas where there are risks of harm to users.

While Bing does not remove such content entirely from the index, Bing applies algorithmic interventions so that users without a clear intent to find such information are protected from accidentally being presented with harmful or misleading information.

Bing monitors evolving trends and threats to continually add to the list of high-risk queries and conversation prompts that receive this higher level of mitigation measures. The monitoring process includes daily pipelines that monitor fact checking trends, tracking known low authority sites to identify new high-risk topics, and other manual intelligence gathering. For any identified high-risk queries, Bing data mines similar sites, queries, and conversations in order to capture relevant permutations and similar high-risk queries.

Content moderation and incident response

Microsoft respects freedom of expression. At the same time, in accordance with Microsoft policies and principles around responsible AI, privacy, digital safety, information integrity, and other critical issues, Bing has developed a safety system intended to provide a safe search experience for Bing users. If Bing receives requests to remove content from individuals, businesses, and governments, in limited cases, where quality, safety, user demand, relevant laws, and/or public policy concerns exist, Bing might remove results, inform users of certain risks through public service announcements or warnings, or provide users with options for tailoring their content. Bing limits removal of search results to a narrow set of circumstances and conditions to avoid restricting Bing users' access to relevant information.

Content moderation based on legal demands

Certain countries have laws or regulations that apply to search service providers and require search engines to remove links to certain indexed pages from search results. Some of these laws allow specific individuals or entities to demand removal of results (such as for copyright infringement, libel, defamation, personally identifiable information, hate speech, or other personal rights), while others are administered and enforced by local governments.

When Microsoft receives a legal request or demand, Bing aims to balance its support for freedom of expression and for free access to relevant content with compliance with local laws. Bing reviews and assesses the request or demand, including the reason and basis for it, the authority or rights of the requesting party, Bing's applicable policies, and Bing's commitments to its users with regard to freedom of expression, human rights, and freedom of information. Bing then determines whether and to what extent access to the content should be removed and takes action. Examples of legal request areas with global approaches are included below.

Child sexual exploitation and abuse imagery (CSEAI) and related materials

The production, distribution, and access to CSEAI materials is universally condemned as a major societal harm and is generally illegal in most jurisdictions. Microsoft policy strictly prohibits these activities on its services, Bing works with third-party technology and industry groups, law enforcement, and governmental and non-governmental organizations to help stop the spread of this horrific content online.

Bing proactively removes illegal CSEAI content from entering the search index. Bing takes a more proactive approach to tackling this issue by preventing pages from entering the index that have been reviewed by credible agencies or identified using Microsoft PhotoDNA hash-matching technology and found to contain or relate to the sexual exploitation or abuse of minors, and regularly reviews content in the index for newly identified CSEAI.

Bing removes pages from its index that have been identified by the Internet Watch Foundation (IWF) in the UK, National Center for Missing and Exploited Children (NCMEC) in the US, and Freiwillige Selbstkontrolle Multimedia-Diensteanbieter (FSM) in Germany, in their good faith judgment, as hosts or providers of access to child sexual abuse material. Removing these links from displayed search results does not block the materials from being accessed on the web or discovered through means other than Bing, but it does reduce the ability of those who would seek it out or profit from it by removing it from the Bing Search index.

The Bing Visual Search feature allows users to use an image as a query to search for similar images. Bing also uses the hash-matching technologies PhotoDNA and MD5 to detect matches of previously identified CSEAI in these images provided by users. In the context of the immediate search, the use of these technologies furthers Bing's goal to avoid inadvertently surfacing potentially harmful web content to users. More broadly, images uploaded to Bing Visual search typically contribute to training Bing's image-matching algorithms; by scanning images Bing helps to ensure that CSEAI is not included in training data.

Copyright infringement

Bing encourages respect for Intellectual Property (IP) rights, including copyrights, while also recognizing the rights of users to engage in uses that may be permissible under applicable copyright laws. If Bing receives a legally sufficient notice of copyright infringement from the copyright owner or its authorized agent, Bing may remove from its search results links to webpages containing material infringing the rights of the owner of copyrighted content.

If a rights holder has an IP concern about a website's content that is linked to by Bing, or a Bing ad, they can review and contact Bing through the Report Infringement page.

Content moderation based on quality, safety, and user demand

In certain circumstances relating to quality, safety, and user value, Bing may decide to remove certain results or may warn or educate users or provide options for tailoring results.

Spam

Certain pages captured in the Bing index may turn out to be pages of little or no value to users and/or can have characteristics that artificially manipulate the search in order to distort their relevance relative to pages that offer more relevant information. Some of these pages include advertisements and/or links to other websites that contain mostly ads, and no, or solely superficial, content relevant to the subject of the search. To improve the search experience for users and deliver more relevant content, Bing might remove such search results or adjust Bing algorithms to prioritize more useful and relevant pages in search results. Bing's general abuse/spam policies, detailed in Bing's Webmaster Guidelines, include details regarding prohibiting certain practices intended to manipulate or deceive the Bing search algorithms.

Sensitive personal information, including nonconsensual distribution of intimate images

From time to time, webpages that are publicly available may intentionally or inadvertently contain sensitive personal information posted without the consent of the individual identified or in circumstances that create security or privacy risks. Examples include inadvertent posting of private records, private phone numbers, identification numbers and the like, or intentionally and maliciously posting email passwords, login credentials, credit card numbers, or other data intended to be used for fraud or hacking. Upon verification, Bing will remove such search results.

Another example is when someone shares adult or sexually explicit images of another person online—whether real or an inauthentic image that is created using AI or other photo editing tools—without that person's consent. Bing may remove links to such photos and videos from search results when it is reported and verified. To report unauthorized online photos and videos, including deepfakes, victims can complete a form on Bing's reporting web page, Report a Concern.

As noted in the reporting form, it is important to remember that information will remain available on the original website even if Bing has removed the link from search results. The website owner is in a position to remove the content from its website.

Adult content

Bing endeavors to avoid delivering content that can be offensive or harmful when it was not requested and provides SafeSearch settings to allow users (and for those using Family Safety features, other users in their "family" account) to control what type of adult content may appear in search results. "Strict" mode prevents adult text and images, "moderate" (the default setting in most countries) restricts explicit images, and the "off" mode allows any manner of content to be displayed in search results. By default, in most countries or regions, SafeSearch is set to Moderate, which restricts visually explicit search results but does not restrict explicit text. Users can choose to change their SafeSearch settings at any time.

Different countries or regions may have different local customs, religious or cultural norms, or local laws regarding the display of adult content or search results accessing adult content. This may affect default SafeSearch settings for Bing in some countries.

Internet PSAs and warnings

In some cases, Bing provides information to the users to help educate them about the potential risks associated with third-party content appearing in the search results, using tools such as PSAs or warnings. For example, Bing provides links to suicide prevention resources when a user's query expresses a possible suicide intent. Microsoft Bing also provides warning notices on certain URLs appearing in search results where it has reliable information that the link contains possibly harmful content. Such notices appear on links to sites Bing has determined to contain harmful malware that could damage a user's computer.

Reports and appeals

Reporting problematic content

If a user finds objectionable content in Bing, the user can report it. Bing values and reviews reports and works continuously to improve the search experience.

Copyright rightsholders or their agents may submit notices of copyright infringement and Bing will block infringing URLs from display in Search. Bing provides notices when copyright content has been blocked. When links have been removed from a page, Bing policy is to note on the search results page that "Some results have been removed." Bing offers webmasters who have created Bing Webmaster accounts to appeal decisions regarding content downranking or removal. Webmasters may appeal such decisions via Bing Webmaster Tools.

In the search engine results page, Bing provides notice to its users when Bing blocks access to a URL that would have otherwise appeared for a particular search query: "Some results have been removed." Other specialized user notices may be displayed on a case-by-case basis. Bing provides a general summary of when the service intervenes in search results through the How Bing Delivers Search Results.

Bing publishes transparency reports regarding content removal as part of the Microsoft Reports Hub, including Copyright Content Removal Requests Report, Right to be Forgotten Content Removal Requests Report, and Government Requests for Content Removal Report.

Bing's approach to risk on generative AI features

Feature-specific risks and mitigations

The Systemic Risk Assessment process also considered how risk profiles may differ and how systemic risks may manifest across generative AI features uniquely and specialized mitigations intended to address the unique risks presented by generative AI.

Generative AI risk profile

While users can create AI-generated content using generative AI features available on Bing, and users could choose to download or share this content via separate means, this AI-generated content is not indexed by the core search feature nor disseminated through the Bing service. The nature of generative AI features and Bing's non-dissemination of the material is a substantial consideration impacting Bing's risk profile, as one user's generated content is not automatically surfaced to other generative AI or Search users.

Some ways that systemic risks can manifest on generative AI features absent sufficient mitigations include:

- Bad actors can attempt to exploit features to generate content that is unsafe, misleading, fraudulent, or otherwise harmful.
- Preemptively blocking responses from being generated could unnecessarily restrict users' fundamental rights to access to information or freedom of expression.
- Generated content could perpetuate bias, propagate false or misleading information, or contribute to echo chambers.
- Generated content may include content that is unsafe, misleading, fraudulent, private, or otherwise harmful.
- As with traditional web search, links contained in the search index could direct users to third-party websites that contain illegal, unsafe, misleading, fraudulent, private, or otherwise harmful content.
- Bad actors could attempt to breach generative AI security and safety systems to access or expose user prompts or other data retained by the search engine.

User counts

The active audience engaging with generative AI features is significantly lower than the total monthly active user count for core search, which informs the risk prioritization of each feature given the impact of any potential risk is lower compared to those areas with a larger opportunity to manifest risk.

Overall approach to mitigating risks in generative AI features

Copilot in Bing strives to provide diverse and comprehensive responses grounded in search results with a commitment to free and open access to information. At the same time, Bing's product quality efforts include working to avoid inadvertently promoting potentially harmful content to users.

Like other transformational technologies, harnessing the benefits of AI is not risk-free. A core part of Microsoft's Responsible AI (RAI) program and Standard is the requirement to identify potential risks, measure their propensity to occur, and build mitigations to address them.

Copilot in Bing benefits from the existing Bing defensive and algorithmic safety measures, and the ranking algorithms and content policies described throughout this report continue to be a primary defense against manipulation and abuse, supplemented by interventions designed specifically to address risks for generative AI features.

Content provenance and watermark

Bing invests in helping users identify AI-generated content. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark, imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The watermark complements the content provenance technology to help ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the Coalition for Content Provenance and Authenticity (C2PA).

Detailed descriptions of generative AI mitigations

Responsible AI program

All Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's Responsible AI Standard and undertake responsible AI reviews to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch. Copilot in Bing and Image Creator from Bing were developed in accordance with Microsoft's AI Principles and Microsoft's Responsible AI Standard. Microsoft's RAI program requires Microsoft personnel, before launching or updating an AI feature or service, to follow certain steps in order to *map*, *measure*, and *mitigate* the risks of harm that the update, feature, or service may present. Part of this process is the development and application of RAI metrics. Mapping the risks of harms, measuring the extent of those harms, and then mitigating those harms is a cornerstone of Microsoft's approach to RAI. More detail on how Microsoft approached the development of Copilot with safety in mind is available on Copilot in Bing: Our approach to Responsible AI.

Risk identification

Bing has conducted extensive red team testing in collaboration with its third-party model licensor, OpenAI. This testing is designed to assess how the latest technology works without any additional safeguards applied to it. The specific intention of the red team testing is to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. The combined learnings across OpenAI and Microsoft contribute to advances in model development, and - for Microsoft - it informs Microsoft's understanding of risks and mitigation strategies for Copilot in Bing.

In addition to model-level red team testing, a multidisciplinary team of experts have conducted numerous rounds of application-level red team testing on Copilot in Bing. This process helped better understand how the system could be exploited by adversarial actors and improve Bing's mitigations. Non-adversarial stress-testers also extensively evaluated new features for shortcomings and vulnerabilities. Post-release, the new AI experiences are integrated into the engineering organization's existing production measurement and testing infrastructure. For example, red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that the team uses for improving the system.

Measurement

Red team testing and stress-testing can surface instances of specific risks, but in production users will have millions of different kinds of conversations with Copilot in Bing. Moreover, conversations are multi-turned and contextual and identifying harmful content within a conversation is a complex task. To better understand and address the potential for risks in Copilot in Bing AI experiences, Bing developed additional RAI metrics specific to those AI experiences for measuring potential risks. These metrics include jailbreaks, harmful content, and ungrounded content. The team also enabled measurement at scale through partially automated measurement pipelines. Each time the product changes, existing mitigations are updated, or new mitigations are proposed. The team then updates measurement pipelines to assess both product performance and the RAI metrics.

As an illustrative example, the updated partially automated measurement pipeline for harmful content includes two major innovations: conversation simulation and automated, human-verified conversation annotation. First, responsible AI experts built templates to capture the structure and content of conversations that could result in different types of harmful content. These templates were then given to a conversational agent which interacted as a hypothetical user with Copilot in Bing, generating simulated conversations. To identify whether these simulated conversations contained harmful content, Bing took guidelines that are typically used by expert linguists to label data and modified them for use by GPT-4 to label conversations at scale, refining the guidelines until there was significant agreement between model-labeled conversations and human-labeled conversations. Finally, Bing used the model-labeled conversations to calculate an RAI metric that captures the effectiveness of Copilot in Bing at mitigating harmful content.

Bing's measurement pipelines enable the service to rapidly perform measurement for potential risks at scale. As Bing identifies new issues through the preview period and ongoing red team testing, it continues to expand the measurement sets to assess additional risks.

Mitigations

As Bing identified potential risks and misuse through processes like red team testing and stress-testing and measured them with the innovative approaches described above, it developed additional mitigations to those used for traditional search. Below, Bing describes some of those mitigations. Bing continues to monitor the Copilot in Bing AI experiences to improve product performance and mitigations. While any of the mitigations work to support safety for Copilot in Bing, not all of the mitigations were undertaken by Microsoft. For example, the generative model behind Copilot in Bing was licensed from OpenAI, who performed Reinforcement Learning from Human Feedback ("RLHF") and data filtering.

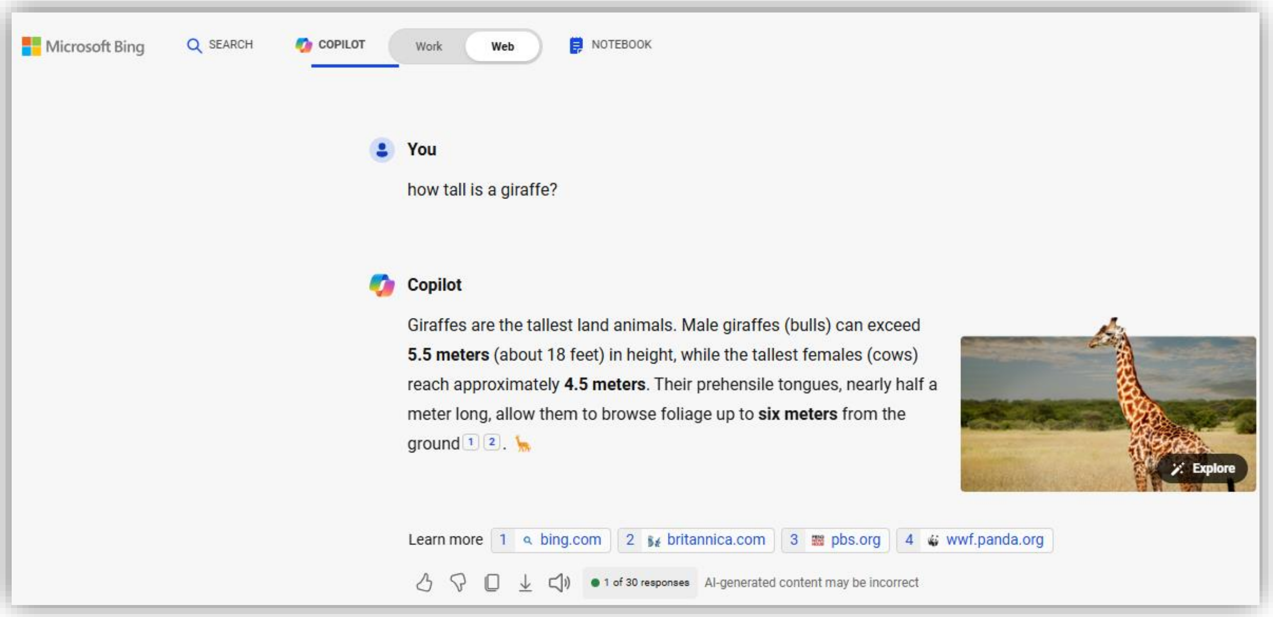
Grounding in search results

Copilot in Bing is designed to provide responses supported by the information in web search results when users are seeking factual information. For example, the system is provided with text from the top search results and instructions via the metaprompt to ground its response.

In summarizing content from the web, Copilot in Bing may include information in its response that is not present in its input sources. In other words, it may produce ungrounded results (these are often called "hallucinations"). The early evaluations have indicated that ungrounded results in Copilot in Bing may be more prevalent for certain types of prompts or topics than others, such as asking for mathematical calculations, financial or market information (for example, company earnings, stock performance data), and information like precise dates of events or specific prices of items. This does not apply to creative including image generation prompts.

Microsoft has taken several measures to mitigate the risk that users may over-rely on ungrounded generated content in summarization scenarios and chat experiences. For example, responses in Copilot in Bing that are based on search results include references to the source websites so that users can verify the response and learn more. Users are also provided with explicit notice that that they are interacting with an AI system and are advised to check the web result source materials to help them use their best judgment.

Figure 14: Copilot in Bing disclaimer screenshot



You

how tall is a giraffe?

Copilot

Giraffes are the tallest land animals. Male giraffes (bulls) can exceed **5.5 meters** (about 18 feet) in height, while the tallest females (cows) reach approximately **4.5 meters**. Their prehensile tongues, nearly half a meter long, allow them to browse foliage up to **six meters** from the ground[1][2]. 🦒

Learn more 1 bing.com 2 britannica.com 3 pbs.org 4 wwf.panda.org

● 1 of 30 responses AI-generated content may be incorrect

Defensive search support

Core defensive search interventions generally carry through to Copilot in Bing, such as by ensuring that Copilot in Bing responses in areas prone to attacks or data voids are grounded in web results that reflect an additional authority signal boost in search ranking.

AI-based classifiers and metaprompting to mitigate potential risks or misuse

The use or misuse of LLMs may produce content with potential systemic risk impact when prompted. For example, users may prompt the system to produce content related to self-harm, violence, or hate speech. Bing has developed prompt classifiers, content filters, and metaprompting on top of the service's pre-existing, robust AI-based classifiers, to help reduce the risk of returning these types of content.

Metaprompting involves giving instructions to the model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. For example, the metaprompt may include a line such as "communicate in the user's language of choice." Taken together, classifiers and metaprompts may function like this: a classifier identifies and flags a prompt as requesting hate speech and then a metaprompt instructs Copilot in Bing to decline to respond, as depicted in the figure below.

Metaprompts

Copilot in Bing and Image Creator from Bing deploy metaprompts are safe and align with Microsoft's RAI principles and user expectations. Metaprompting involves giving instructions to the model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. For example, the metaprompt may include a line such as "communicate in the user's language of choice." Metaprompts that Microsoft has adopted to mitigate the risks of a range of potential harms are designed to:

- reduce the risk that the output contains harmful content, including content that is violent, sexual, demeaning, or otherwise problematic;
- avoid the generation of content that may infringe on third-party IP rights; or
- avoid disclosure of information that might enable users to jailbreak other safety systems that Microsoft has implemented.

Classifiers and filters

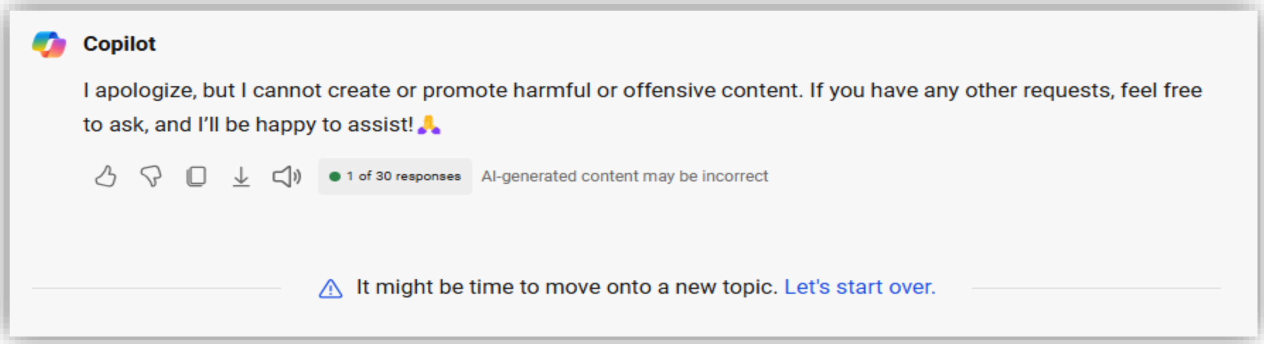
Copilot in Bing and Image Creator from Bing deploy classifiers to analyze and "classify" user inputs and generated responses in order to assess the level of risk that they might include (or generate) harmful content. Classifiers classify text to "flag" different types of potentially harmful content in search queries, chat prompts, or generated responses. Microsoft uses AI-based classifiers and content filters, which apply to all search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by new generative AI features such as Copilot in Bing. Flags lead to potential mitigations, such as restricting responses to the user, diverting the user to a different topic, or redirecting the user to traditional search. Microsoft has also implemented additional filtering and classifiers to prevent Copilot in Bing chat responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

Blocklists

Both Copilot in Bing and Image Creator from Bing use blocklists to restrict the generation of content for user inputs that contain certain names, phrases, or terms as an additional way to prevent generation of potentially harmful content. Microsoft continues to invest in improving the comprehensiveness and efficacy of these blocklists.

Taken together, classifiers and metaprompts may function like this: a classifier identifies and flags a prompt as requesting hate speech and then a metaprompt instructs Copilot in Bing to decline to respond, as depicted in the figure below.

Figure 15: Copilot in Bing response to request to produce hate speech screenshot



Copilot

I apologize, but I cannot create or promote harmful or offensive content. If you have any other requests, feel free to ask, and I'll be happy to assist! 🙏

⚠ It might be time to move onto a new topic. Let's start over.

Protecting privacy in Visual Search in Copilot in Bing

When users upload an image as a part of their chat prompt (e.g., to seek information on content in an image), Copilot in Bing will employ face-blurring technology before sending the image to the AI model. Face-blurring is used to protect the privacy of individuals in the image. The face-blurring technology relies on context clues to determine where to blur and will attempt to blur faces. With the faces blurred, the AI model may compare the input image with those of publicly available images on the Internet. As a result, for example, Copilot in Bing may be able to identify a famous basketball player from a photo of that player on a basketball court by creating a numerical representation that reflects characteristics such as the player's jersey number, jersey color, and the presence of a basketball hoop. Copilot in Bing does not store numerical representations of people from uploaded images and does not share them with third parties. Copilot in Bing uses numerical representations of the images that users upload for the purpose of responding to users' prompts, then they are deleted within a period of time after the chat ends.

If the user asks Copilot in Bing for information about an uploaded image, chat responses may reflect the impact of face-blurring on the model's ability to provide information about the uploaded image. For example, Copilot in Bing may describe someone as having a blurred face.

Limiting conversational drift

Microsoft learned that very long chat sessions can result in responses that are repetitive, unhelpful, or inconsistent with Copilot in Bing's intended tone. To address this conversational drift, Bing limited the number of turns (exchanges which contain both a user question and a reply from Copilot in Bing) per chat session. The team continues to evaluate additional approaches to mitigate this issue.

Prompt enrichment

In some cases, a user's prompt may be ambiguous. When this happens, Copilot in Bing may use the LLM to build out more details in the prompt to help ensure users get the response they are seeking. Such prompt enrichment does not rely on any knowledge of the user or their prior searches, but instead on the AI model. These revised queries will be visible in the user's chat history and, like other searches, can be deleted using in-product controls.

As an example of this, a user may enter a prompt for "Switzerland travel" and Copilot in Bing may enrich the prompt with context such as inferring that the user wants travel tips, querying for landmarks, attractions, and destinations in Switzerland. Then Copilot in Bing could respond with "Switzerland has a lot to offer. Here are some highlights and tips for your trip: Landmarks and attractions include Matterhorn (an iconic mountain peak), Jungfraujoch (known as the 'Top of Europe." Copilot in Bing may title the conversation in the chat history as "Switzerland Travel Tips"

User-centered design and user experience interventions

User-centered design and user experiences are an essential aspect of Microsoft's approach to responsible AI. The goal is to root product design in the needs and expectations of users. As users interact with Copilot in Bing for the first time, Copilot in Bing offers various touchpoints designed to help them understand the capabilities of the system, discloses to them that Copilot in Bing is powered by AI, and communicates limitations. The experience is designed in this way to help users get the most out of Copilot in Bing and minimize the risk of overreliance.

Elements of the experience also help users better understand Copilot in Bing and their interactions with it. These include chat suggestions specific to RAI (for example, how does Bing use AI? Why won't Copilot in Bing respond on some topics?), explanations of limitations, ways users can learn more about how the system works and report feedback, and easily navigable references that appear in responses to show users the results and pages in which responses are grounded.

Terms of Use and Code of Conduct

Copilot AI Experiences Terms govern the use of Copilot in Bing. Image Creator Terms govern the use of Image Creator from Bing. Users must abide by the terms, which inform them of permissible and impermissible uses (including prohibited content and conduct) and the potential consequences of violating terms, including account suspension and appeal mechanisms. The terms also provide additional disclosures for users and serve as a reference for users to learn about Copilot in Bing or Image Creator from Bing.

Operations and rapid response

Copilot in Bing's ongoing monitoring and operational processes are used to address when Copilot in Bing receives signals, or receives a report, indicating possible misuse or violations of the Terms of Use or Code of Conduct. Content that may involve imminent harm is the highest priority. The second priority is key priority policy areas (e.g., abusive AI content in elections), and the third priority is significant external visibility of issue. Additionally, other considerations such as the content's language, the region from where the content originated, and the media type also affect the prioritization of the review process. Most content issues are resolved within 24 hours regardless of prioritization.

Feedback, monitoring, and oversight

The Copilot in Bing and Image Creator from Bing experiences build on existing tooling that allows users to submit feedback and report concerns, which are reviewed by Microsoft's operations teams. Microsoft's operational processes have also expanded to accommodate the features within Copilot in Bing experience, for example, updating the Report a Concern page to include the new types of content that users generate with the help of the model. Bing's approach to identifying, measuring, and mitigating risks will continue to evolve as it learns more.

Automated content detection - CSEAI

When users upload images as part of their chat prompt, Copilot in Bing deploys tools to detect CSEAI, most notably PhotoDNA and MD5 to detect matches of previously identified CSEAI in these images provided by users. In the context of the immediate search, the use of these technologies furthers Bing's goal to avoid inadvertently surfacing potentially harmful web content to users. Microsoft reports CSEAI detected on its services to the NCMEC, as required by U.S. law.

Automated content detection – Other

When users upload files to analyze or process, Copilot in Bing deploys additional automated scanning to detect content that could lead to risks or misuse, such as text that could relate to illegal activities or malicious code.

Bing's approach to risk on ancillary search features

Feature-specific risks and mitigations

The Systemic Risk Assessment process considered how the risk profile may differ and how systemic risks may manifest across ancillary search features uniquely as well as what feature-specific mitigations are in place to address these risks.

Risk profile for ancillary search features

Some ways that systemic risks can manifest on ancillary search features absent sufficient mitigations include:

Maps

- Maps business listings may include illegal, fraudulent, misleading, or otherwise harmful locations.
- Street-level imaging features may include private or sensitive images.

- Maps may recommend fraudulent, illegal, or otherwise harmful locations to users of any age.
- Maps boundaries and names may not sufficiently take into consideration cultural and linguistic factors.

Shopping

- Shopping may include listings for illegal, fraudulent, discriminatory, or otherwise harmful products.

News

- News results may contain content from or links to third party news providers and articles could potentially contain inaccurate, biased, misleading, or low authority content
- News results may contain news content that includes misinformation, discriminatory themes, scams, violence, or other harmful substances.
- News source selection for recommendation may limit users' access to pluralism of information.

Travel

- Travel Stories may contain illegal, violent, discriminatory, private, fraudulent, or otherwise harmful content.
- Travel results or recommendations may include scams or fraudulent booking opportunities.

Real Estate

- Fraudulent real estate listings may appear on Bing Real Estate.
- Real Estate recommendations could perpetuate bias, stereotypes, or inequalities.
- Real estate listings could include discriminatory, illegal, or otherwise harmful content.
- Real estate messages could include harassing, discriminatory, illegal, or otherwise harmful content.
- Real estate listings could lead users to unsafe off-platform interactions.

Advertisements on Bing

- Advertising content may include illegal, fraudulent, misleading, or otherwise harmful content.
- Advertisements may link to malware or fraudulent, low authority, or otherwise harmful sites.
- Advertisements may target users of any age in a way that is discriminatory or negatively impacts their right to privacy or protection of personal data.

User counts

We consider these ancillary search features because the active audience engaging with these features is significantly lower than the total monthly active user count for core search, which informs the risk prioritization of each feature given the impact of any potential risk is lower compared to those areas with a larger opportunity to manifest risk.

Approach to mitigating risks in ancillary search features

In addition to the broad mitigations described in the prior segments of this Report, Bing has implemented a variety of mitigations specific to ancillary search features highlighted here:

Maps

Maps has implemented and continues to refine a variety of risk mitigation measures, including:

- Maps content is primarily licensed from industry-leading partners.

- Bing reviews images uploaded to Maps locations and locations suggested for addition to Maps by users or businesses prior to posting.
- Maps honors user and government requests to blur images.
- Maps follows requirements for cultural sensitivity or laws to display differing names or geopolitical boundaries depending on the location of the individual conducting the search.
- Bing enables users to report concerns related to Maps locations or content.

Shopping

- Shopping relies on a two-level review to vet products. First, it relies on the Ads team to provide a listing of products reviewed for compliance with Bing’s Advertising Policies. This will filter the vast majority of products. At that point Shopping conducts its own vetting process of products from Shopping crawl data, with additional requirements for what content can be recommended to users.
- Bing also employs proactive blocking using classifiers to detect offensive text within images and product names on Shopping.
- Bing monitors the presence of offensive text by utilizing automated labeling and generates reports accordingly. If any anomalies are identified, a review team meets to determine the proper course of action.
- Similar to other ads on Bing, Bing Shopping allows users to report potentially harmful listings and addresses these through reactive takedown measures.

News

- Bing scans news content using a robust set of classifiers to identify potentially harmful content.
- Microsoft also supports reputable third-party credibility ratings of online news, and Bing has been a leader in the use of trustworthiness indicators in search to help users evaluate the credibility of news sources they encounter. This includes reliance on signals of page authority and information regarding misinformation narratives that Bing receives from various third-party partners, such as NewsGuard, the Global Disinformation Index (GDI), and the Reporters sans Frontières (RSF) Journalism Trust Index.
- Recommended News content is generally limited to top tier news sources. Bing identifies top tier sources in each country using in-market teams to conduct in-depth editorial reviews.
- Bing enables users to report concerns on News content and addresses these concerns accordingly.

Travel

- Bing directs users to booking options from vetted industry-leading partners.
- Travel Stories are reviewed for potentially harmful content.
- Bing conducts additional quality assurance checks on Travel Stories.
- Bing enables users to report any concerning content on Travel Stories or other content.

Real Estate

- Bing actively scans rental listing descriptions and images to ensure compliance with its content policies, employing automated tools to detect inappropriate content, hate speech, or offensive images.
- Bing has implemented sophisticated fraud detection models that proactively identify fraudulent rental listings by checking various signals, such as price inconsistencies across platforms and the

legitimacy of Microsoft Accounts. This includes limiting the number of listings per Microsoft account to combat spam and coordinated inauthentic activities.
- Bing maintains dashboards to monitor the frequency of flagged, demoted, and removed content, enabling the identification of emerging trends and issues in real time.
- Bing allows users to report a concern with any listing on Bing Real Estate.
- Bing provides the ability for users to block any other user from sending them messages.

Advertisements on Bing

Microsoft has a robust Advertising review and compliance process that is run wholly separately from Bing. This process includes the following mitigations:

- Microsoft Advertising hosts and manages content provided by its advertisers, pursuant to the Microsoft Advertising Agreement, which includes the Microsoft Advertising Network Policies. Microsoft Advertising has clear and regularly enforced content policies and practices that prevent advertisements negatively impacting user safety. The Microsoft Advertising Policies set out the requirements for ad content, including criteria upon which ad content will be removed. Microsoft requires its advertisers and partners to comply with its policies throughout their use of the Microsoft services. Microsoft Advertising also has a set of Relevance and Quality Policies to manage the relevance and quality of the advertisements that it serves through its advertising network.
- Bing ensures that ads meet high relevance and quality standards by leveraging the Microsoft Advertising policies and procedures, which include both manual and automated processes to filter out and mitigate harmful online traffic, including fraud and malware.
- Advertisers retain ownership of and responsibility for their ad content, but the advertiser must agree to the terms when signing up for a Microsoft Advertising account.
- Advertisements are reviewed for compliance with the Microsoft Network Advertising policy prior to being posted. This review leverages machine-learning techniques, automated screening, the expertise of its operations team, and dedicated user safety experts.
- Microsoft applies robust classifiers to scan ads and websites promoted by advertisers for content policy compliance, misleading behavior, malware, phishing, and fraudulent links.
- Users can report ads that they believe violate compliance associated with a particular harm through the feedback bubble found on the site or using the Report a Concern form. Users will receive notification after action is taken.
- Microsoft Advertising conducts a manual review of advertisements flagged to its customer support team and removes advertisements that violate its policies.
- Microsoft does not allow political advertising within the Microsoft Advertising ecosystem, which supports advertising on Bing. Microsoft Advertising policies 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 and ballot measures also are barred.
- Microsoft Advertising's policies also prohibit certain types of advertisements that might be considered issue-based or might violate local regulations.
- Microsoft Advertising employs a robust filtration system to detect robotic traffic and other harmful cyber-attacks.

- 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 across internal teams to verify data accuracy and integrity.
- Microsoft Advertising is a signatory to the EU Code of Practice on Disinformation (COPD), which requires biannual reporting and compliance with anti-disinformation measures to improve Bing's ability to fight misinformation across the EU, including commitments to providing users with indicators of content provenance and fact checks, as well as providing researcher access to data.
- Finally, any ads and associated components within an advertiser's account which violates policies within the following categories (Ad Requirements, Disallowed Content, Extensions, IP, Legal and Privacy, Media Formats, Product Ads, Relevance and Quality, Restricted Content), will be subject to a "three strikes" enforcement penalty before Bing will suspend that user's account use of Microsoft Advertising services. If an advertiser is under a strike penalty, they will not be able to log into their account through Editor or the Mobile app but can access through the Web User Interface (UI). Bing Microsoft Advertising determines strikes by policy categories related to the violation, or any violation that may pose a risk to the safety or security of Bing customers or users. Enforcement actions may include suspension of ads and associated components of the advertiser, limiting creation of new accounts, and suspending the advertiser's account temporarily or permanently.

How Bing monitors effectiveness of risk mitigation efforts

Bing monitors and regularly reviews the efficacy of relevant and key mitigations using metrics to identify additional areas for improvement. Bing also monitors the content moderation and incident response processes to ensure its system is robust and reliable. Bing conducts automated social listening and provides users with channels to provide feedback and report their concerns to better understand user experiences for further product improvement. Additionally, Bing works with internal and external subject matter experts in key policy areas to identify new threat vectors or improved mechanisms to help users engage safely with content in search results (e.g., new approaches to digital literacy messaging).

Measurement metrics

Bing uses a metric known as the Defensive Defect Rate (DDR) as a primary means to measure the efficacy of implemented safety mitigations and the presence of content that violates Bing policies. DDRs measure "defects" which are defined as any time that, in response to user queries, the top results on the search pages show content that is not in accordance with Bing's product principles. Bing measures the service against "adversarial" metrics sets that are designed to mostly contain high-risk queries that are prone to lead to harmful results to measure associated defects. Bing sets objectives to improve these metrics on an annual basis and reviews them on a monthly basis to drive continuous improvement. Once Bing has achieved the set objective, the bar is raised with a new adversarial set or stricter evaluation criteria. This culture of continuous improvement has allowed Bing to keep improving the safety of the service over several years.

Bing also conducts metric analysis based on DSA systemic risks to measure the effectiveness of Bing's ranking algorithms and additional ranking interventions per DSA systemic risk. These metrics were used in the risk assessment as an underlying analysis.

- The metrics may be used to evaluate feature and team performance, and teams are accountable for their performance.
- The metrics are reviewed monthly in shiproom meetings with Bing leadership and feature teams; when regressions are observed, teams conduct issue identification and implement fixes or drive improvements.
- DDR measurement is also a part of pre-launch assessment.
- Bing conducts analysis on DDR for each DSA systemic risk as an underlying analysis for the formal risk assessment.
- Microsoft works to systematically measure the efficacy of Copilot in Bing's grounding to factual information in web search results.

Shiproom review

Every proposed change to Bing, Copilot in Bing, or Image Creator from Bing must go through a launch readiness review, which Microsoft refers to internally as a "shiproom" review. This review process requires product teams wishing to adopt changes to these features to conduct a number of different tests and evaluations and to complete corresponding documentation. This includes both offline testing and A/B testing online. Among other things, the product team must have conducted metrics testing on the proposed change, including DDR, and must provide data demonstrating whether the proposed change would result in improvements, regression, or no change against these metrics, and explain any additional mitigations implemented to address identified issues. The shiproom review also confirms that additional reviews such as for digital safety, security, privacy, and accessibility have been completed prior to launch.

Incident response process testing and Service Level Agreements (SLAs)

Bing has a robust reporting and reactive response infrastructure that allows it to quickly action notices of illegal materials and other policy violating content (including exposed personal information, CSEAI, Nonconsensual Intimate Imagery (NCII), etc.) from governments, users, or other stakeholders. Support teams are trained in applicable requirements and policies and provide escalation paths where needed for complex issues, including to local legal experts. The Bing team also relies on Microsoft's extensive team of subject matter experts and external experts to identify and quickly address emerging concerns.

Bing also has an incident response process for high-risk escalations to be handled urgently. Bing teams are trained to use IcM (an internal incident management tool), and email alerts will be sent to cross-functional teams and Bing leaderships accordingly.

- Both content moderation and incident response processes have specific SLAs as part of the Objectives and Key Results (OKRs), and teams are committed to meeting the SLAs agreed upon.
- Both processes have random testing, to ensure that the end-to-end procedure runs smoothly.

Social listening

Bing maintains social listening pipelines where insights and user feedback on Bing's features are collected from the Internet. These insights and user feedback are manually reviewed, analyzed daily, and shared across Bing's product teams and product engineering leadership in a daily report. These reports inform Bing as to the public concerns, issues, or emerging trends and can serve as barometer of public sentiment on various topics related to Bing.

User feedback and reports

At the bottom of each page on Bing, Copilot in Bing and Image Creator from Bing, users can find a link entitled "Feedback;" this link directs to a text box where users can share their views on search results, generated responses, or other aspects of the services. User feedback is triaged in accordance with Microsoft internal guidelines and used as insights for product development.

When users have specific concerns about information they see on Bing, Copilot in Bing, or Image Creator from Bing, they may also report concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing and in generative AI interfaces. Users can report:

- Exposed personal or private information (e.g., sensitive confidential information like credit card numbers or passwords, personal identifying information, information about minors, fake pornography or unrelated pornographic results, or nonconsensual intimate imagery)
- Intellectual Property Infringement (e.g., copyright or trademark infringement)
- Unlawful content
- Malicious websites or spam (e.g., malware, phishing, spam, or exploitative content removal practices)
- Unexpected offensive or harmful material (e.g., reporting unexpected adult, violent, or gore in results, non-consensual intimate imagery, or CSEAI)
- Any other concerns (e.g., broken links, feature issues, search suggestions, etc.)
- Issues with AI-powered features (e.g., issues with generated content)

Microsoft teams manually review these user reports, triage reports to appropriate team(s) for review, and take action where appropriate.

External feedback

Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as terrorist and violent extremist content, information integrity and misinformation, CSEAI, responsible AI, and AI-specific risks, hate speech, minors and technology, and copyright and trademark infringement, to ensure that its internal policies, practices, and standards are addressing key concerns third-party stakeholders. These engagements can inform processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. Routine external engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Systemic Risk Assessment methodology

Bing's Systemic Risk Assessment methodology was designed using a variety of available scientific and technical insights and included consultation with civil society organizations representing vulnerable users of Bing products and services.

As part of the Systemic Risk Assessment, the Bing Risk Assessment team considered the probability and severity of inherent risks due to the functioning, use, or misuse of Bing's products and services, identified implemented mitigations, and assessed whether those mitigations are effective, reasonable, and proportionate to the identified risks.

The Bing Risk Assessment team considered possible risk factors and influencers, such as the intentional manipulation of the service, the design of recommender systems and content moderation systems, and regional and linguistic aspects. The process that the Bing Risk Assessment team followed to conduct this period's Systemic Risk Assessment is described here along with a description of the enhancements made in the design of this period's assessment. Additional detail, including definitions and rating scales, is included in the Appendix I: Detailed Risk Assessment and Scoring Methodology.

Risk assessment design enhancements

For the conduct of this period's Systemic Risk Assessment, Bing engaged the support of external assessment experts to support its DSA Systemic Risk Assessment. Bing made several enhancements to this year's assessment, including:

- Implementation of a rigorous, quantitative scoring methodology, leveraging industry standard numeric rating scales, to support and inform the calculation of residual risk in order to enable a more objective prioritization and mitigation of those risks;
- Development of a Risk Assessment Workbook to provide supporting information on how risks and relevant mitigations were assessed and to enable and evidence more quantitative scoring;
- Refined risk areas and the inclusion of risk scenarios for increased coverage of key systemic risks as outlined in Article 34 of the DSA;
- Explicit consideration of the manifestation of each relevant risk scenario on individual Bing features;
- Identification and formalization of quantitative and qualitative inputs to the Systemic Risk Assessment;
- Incorporation of internal and external data for a data-informed probability score;
- Establishment of a systems-based analysis of severity (identifying a level of severity based on the scale and remediability of a given risk's impact across the following complex systems: wellbeing, environmental, political, societal, security, and economic) to help establish a more objective calculation of inherent risk;
- Identification of focus areas for the 2024 Systemic Risk Assessment based on global trends and changes to Bing products and services;
- Continued facilitation of internal workshops with Bing experts to gather information on risks and mitigations across specific Bing features and to test the assumptions of the Bing Risk Assessment team; and

- Expanded mapping of each implemented mitigation to the identified risk scenarios to ensure adequate coverage of inherent risk.

Risk assessment process

Risk definition and discovery

The Bing Risk Assessment team identified twelve risk areas aligned with the risks identified in Article 34 of the DSA for consideration of the potential impact of the use or misuse of Bing's products and services.

1. Civic Discourse and Electoral Processes
2. Consumer Protection and Fraud
3. Discrimination and Hate
4. Freedom of Expression, Information, and Pluralism
5. Human Dignity
6. Illegal Content and Activities
7. Mental and Physical Well-being
8. Right to Private and Family Life
9. Personal Data
10. Public Health
11. Public Security
12. Rights and Protections of Minors

The Bing Risk Assessment team collected cross-functional stakeholder input and information on updates to Bing features since August 2023, on the presentation and prevalence of risks across features, and on the implemented mitigations and ongoing monitoring activities across industry-standard best practices both in writing, through stakeholder engagements, and through a review of existing Microsoft documentation and data.

The Bing Risk Assessment team conducted workshops and interviews with key stakeholders across the organization to deep dive into the presentation of risks and unique mitigation measures across features. The Bing Risk Assessment team reviewed inputs from Microsoft's civil society engagements conducted throughout the year and feedback from other external stakeholders to discuss areas of concern and best practices for mitigation.

The Bing Risk Assessment team gathered key internal metrics related to Bing policy enforcement; deconstructed transparency reporting metrics; and reviewed open-source information related to trends, patterns, and potential systemic risk related to use or misuse of Bing services, including Bing's social listening metrics and insights.

The Bing Risk Assessment team also examined authoritative sources and case studies, including Centre on Regulation in Europe (CERRE) and European Commission Guidelines, on the systemic risk areas, collected relevant public policies and publications, and summarized relevant internal controls.

Inputs to the Systemic Risk Assessment:

1. Bing internal stakeholder responses to mitigation questionnaires.
2. Bing internal stakeholder responses to material changes questionnaires.
3. Collection of Bing's published policies and official communications.

4. Bing internal stakeholder summarized mitigations by risk area.
5. Relevant controls from the DSA Control Inventory.
6. Notes from internal consultations.
7. Notes from external consultations.
8. Internal policy enforcement metrics.
9. Transparency Reporting metrics.
10. Open-source data on public discourse related to Bing and systemic risk areas.
11. Guidance from authoritative sources, including the European Commission, CERRE, and other sources such as civil society or other non-governmental organizations, on the severity of systemic risks.

These inputs are described in more detail in Appendix I: Detailed Risk Assessment and Scoring Methodology.

Risk analysis

The Bing Risk Assessment team assessed the probability and severity of each of the twelve systemic risks stemming from use, misuse, or functioning of its products and services, as well as the maturity of risk mitigation measures implemented, to arrive at a prioritization of systemic risk areas with potentially higher levels of residual risk. The assessment incorporated the referenced eleven inputs from the definition and discovery phase to inform scores and implemented a traditional risk assessment equation to aid in prioritization of risk for action.

The Bing Risk Assessment team also considered for each applicable risk area and mitigation whether and how the following factors influence any of the systemic risks: the design of recommender systems and other relevant algorithmic systems; content moderation systems; applicable terms and conditions and their enforcement; systems for selecting and presenting advertisements; data-related practices; linguistic and cultural considerations; intentional manipulation including inauthentic use or automated exploitation; and amplification and potentially rapid and wide dissemination of illegal or violative content.

Probability

The Bing Risk Assessment team identified key risk scenarios associated with each risk area and considered the potential theoretical manifestation of each risk scenario across various features of the Bing service absent safeguards.

The Bing Risk Assessment team employed a data-driven probability assessment, combining insights from internal, external, and open-source data to assess the likelihood of specific risks stemming from use or misuse of Bing considering the vulnerability of the service absent mitigation measures and user demand or likely frequency of attempts to perpetrate each harm on the service. This methodology follows models used in climate and energy impact assessments.

Bing monitors a number of metrics in identifying and measuring risk across the service. Given that each metric has its own strengths and weaknesses, and thus cannot measure all aspects of a particular risk's potential negative impacts to Bing’s services, the Bing Risk Assessment team incorporated several internal and external data points, including social listening data, survey data, DDR, and content moderation metrics, to inform the relative probability of the risk. The Bing Risk Assessment team highlighted DDR in the individual risk analysis sections below, as Bing teams track and monitor DDR as one of the main metrics to measure overall product health and evaluate the effectiveness of Bing's multi-layered safety system.

- DDR measures the likelihood that certain types of queries might result in harmful content showing in the top search results.
- Bing uses "adversarial sets," sampling queries that are focused on high risk topics to help zoom in on the areas prone to showing harmful content and help better identify gaps and drive targeted improvements. Due to the nature of the "adversarial sets," DDR is not representative of all the traffic but rather focuses on a very small portion of the high risk queries.
- Bing uses a "diluted" DDR to represent the raw DDR of DSA systemic risks across overall search traffic, while also considering the trigger rate of the defensive classifiers. The Defensive classifier trigger rate is <1%, meaning that across all user queries, less than 1% of queries might potentially lead to problematic content.

Severity

The Bing Risk Assessment team leveraged a systems-based assessment of severity to achieve a more objective severity score. This methodology follows models used in environmental impact assessments and entails examining the complex systems impacted by the systemic risk area (geographic, political, security, environmental, societal, and wellbeing), while considering the scale (individual to global) and gravity (or remediability) of impact to arrive at a severity score.

Risks with irremediable impact on a greater number of systems on a broader scale will receive the highest severity scores. For example, risks related to Consumer Protection and Fraud are rated as High (not Critical) severity because, while some impacts may be irremediable, the greatest impact is at the individual level and primarily on economic systems. Risks related to Public Security are rated as Critical as this risk category includes a number of risks with irremediable impacts (loss of life, for example) across multiple systems (security, political, and wellbeing) that reverberate across the local, country, and potentially regional levels.

Maturity of mitigations

Consistent with Bing's Year One risk assessment, the Bing Risk Assessment team organized existing mitigations and controls according to industry-standard best practices and evaluated their implementation according to the DTSP Maturity Rating Scale. The DTSP Maturity Rating scale is a best practice framework that uses a scale from "Ad Hoc" to "Optimized." The Bing Risk Assessment team used this scale to support the evaluation of the reasonableness, proportionality, and effectiveness of Bing's existing mitigations to the probability and severity of the considered risks and risk manifestations and identified potential areas for enhancements to existing mitigations or new mitigations to further reduce the potential impact of the considered systemic risks.

Scoring

The Bing Risk Assessment team calculated the inherent risk for each risk area resulting from combining the ratings of probability and severity. The team then reduced the inherent risk rating by a percentage proportional to the assessed strength of the relevant mitigations. This methodology produces a view of residual risk (i.e., the remaining risk after accounting for applicable mitigations) across the risk areas to enable a more objective measurement of risk and to identify risks or mitigations that may warrant further investment, in alignment with Article 35 of the DSA.

For a more detailed description of the risk calculation, please refer to Appendix I: Detailed Risk Assessment and Scoring Methodology.

Assessment tools

Risk assessment workbook

The Bing Risk Assessment team employed a Systemic Risk Assessment workbook this year to facilitate the standardized completion of Systemic Risk Assessment and to store relevant ratings, scores, and rationale. The Bing Risk Assessment team used this workbook to collect and assess the relevant risk factors and risk scenarios, to identify potential manifestation of risks across Bing products and services absent safeguards, to capture and align implemented mitigations, to map mitigations to identified risks, and to calculate and display risk assessment scores.

Risk assessment results

This section of the Report outlines the Residual Risk remaining within each Risk Area after the application of Bing's implemented mitigations to the identified Inherent Risks. For each Risk Area, the Bing Risk Assessment team has defined the risk areas and included theoretical ways they may manifest on the service, described the calculation of and rationale for the Inherent Probability and Severity scores, and included the key mitigations relevant to that Risk Area. The rating scales and definitions are included in the Appendix I: Detailed Risk Assessment and Scoring Methodology section of the Report.

The Risk Areas are presented in this section in the order of highest Residual Risk, followed by highest Inherent Probability, and then by lowest Maturity of Mitigations.

Figure 16: 2024 DSA Systemic Risk Assessment Ratings

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Human Dignity	Highly Likely	Critical	Critical	Managed	Moderate
Public Security	Highly Likely	Critical	Critical	Managed	Moderate
Mental and Physical Wellbeing	Likely	Critical	High	Defined	Moderate
Rights and Protection of Minors	Likely	Critical	High	Defined	Moderate
Consumer Protection and Fraud	Highly Likely	High	High	Managed	Low
Protection of Personal Data	Highly Likely	High	High	Optimized	Low
Illegal Content and Activities	Likely	Critical	High	Managed	Low
Private and Family Life	Likely	High	High	Managed	Low
Discrimination and Hate	Likely	High	High	Managed	Low
Civic Discourse and Electoral Processes	Likely	Critical	High	Optimized	Low
Public Health	Not Likely	Critical	Moderate	Managed	Low
Freedom of Expression and Information	Not Likely	High	Moderate	Optimized	Low

Areas of Moderate Residual Risk

Considering the application of mitigations to the identified inherent risks, Bing assessed that a majority of the assessed risks fell into a Residual Risk category of Low with four Risk Areas receiving a Moderate risk rating: Human Dignity, Public Security, Mental and Physical Wellbeing, and the Rights and Protection of Minors.

Human Dignity

Risks related to Human Dignity include exposure to content pertaining to sexual exploitation, human trafficking, vulgarity, and gore. Bing's ranking algorithms are designed to protect users from being unexpectedly exposed to content that could negatively affect fundamental rights to human dignity. Bing provides users with a SafeSearch option to control their search experience to filter out adult content and/or content including explicit language. Generative AI features deploy additional mitigations to prevent harmful AI-generated outputs such as pornographic content or graphic violence. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Human Dignity, the Bing Risk Assessment team has assessed the Residual Risk related to Human Dignity on the service to be Moderate.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Human Dignity	Highly Likely	Critical	Critical	Managed	Moderate

Risk Definition

Risk Area	Human Dignity
Risk Definition	Risk that content or activities with actual or foreseeable negative effects on the fundamental rights to human dignity occur on the service.
Theoretical Risk Manifestations absent Mitigations	• Risk that Search results unexpectedly return content that is sexually explicit, vulgar, or violent. • Risk that Bing results include links to sites facilitating or depicting human trafficking. • Risk that Image Creator from Bing is used to create materials that depict sexual exploitation, pornography, or extreme pornography. • Risk that sexually explicit, violent, or vulgar ads appear on Bing. • Risk that News results include content that is sexually exploitative, pornographic, violent, or vulgar. • Risk that Image Creator from Bing is used to create violent, gory, or profane content. • Risk that Copilot in Bing provides responses that are vulgar or include information that facilitates sexual exploitation or human trafficking.

Risk Analysis

The **Inherent Probability** of systemic risks related to Human Dignity stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Highly Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a "Highly Likely" level of probability relative to other potential systemic risks on Bing.
- User intent is generally required for users to access potentially harmful content on Bing. Due to the nature of Bing as a search engine, Bing's goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less authoritative, while working to ensure that users are not misled by such search results or inadvertently exposed to material they

could find harmful or offensive. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.

- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the Inherent Probability and ways that users may be exposed to potentially harmful content vary between core search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** of impact for content and activities that negatively impact Human Dignity is rated as "Critical" primarily due to the gravity of risks within this category, considering the potential for significant harm to wellbeing, societal, economic, and security systems at the individual and potentially broader level. Some risks within this category, such as human trafficking and prostitution, are potentially irremediable. Other risks, such as gore, vulgarity, profanity, or pornographic content, may be remediable but can be significant.

With the rating for Inherent Probability as "Highly Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with negative impacts to Human Dignity on the Bing service is "Critical."

Risk Mitigation

Bing has implemented a robust set of mitigations to address risks related to Human Dignity, including those described in Bing's approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and those summarized below. The mitigations Bing has implemented relative to Human Dignity follow best practices with defined, documented, and managed processes. As such, the **Maturity of Mitigation** efforts Bing has applied to the Human Dignity risk area is assessed as "Managed," which brings the **Residual Risk** rating for down to "Moderate."

While not a proxy for Residual Risk, Bing has measured the DDR for Human Dignity on Bing Search as an average of .81% from April to June 2024. This means that Bing estimates .81% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Human Dignity.

Considering the Moderate Residual Risk, Bing continues to prioritize investment in developing and enhancing strategies to further mitigate and manage risks related to Human Dignity. The key mitigations currently implemented are described below.

Product Development

User reporting of suggestions: To uphold the principle of human dignity, Bing has implemented measures aimed at curbing the generation of suggestions that might tarnish this fundamental right. This includes creating avenues for users to effortlessly report suggestions they deem inappropriate, helping to ensure such concerns are promptly addressed. Through these actions, Bing demonstrates its commitment to fostering a respectful digital space where users feel valued and protected.

Algorithmic prioritization of high authority content: How Bing Delivers Search Results and the Bing Webmaster Guidelines contain Bing's principles for ranking and moderation of third-party content in web results and provide detailed information on the removal of content that violates laws or Bing principles. Bing's investments both in time and resources dedicated to ensuring that its algorithms provide high quality content and avoid returning low quality or harmful content to its users extends to Human Dignity. Content considered to be offensive to human dignity, which includes pages calling for human trafficking, sexual exploitation, pornographic content, violence, gore, vulgarity, and profanity are generally considered and marked as low quality, affecting their visibility in search results.

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undergo an RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot in Bing provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing, and Image Creator from Bing and their features are required to undergo pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

New Feature Launch: Bing relies on Microsoft's extensive cross-company compliance infrastructure to proactively review new products and features to ensure they meet Microsoft's policy commitments in key areas such as privacy, accessibility, digital safety, and RAI.

Product Governance

Code of Conduct: Policies governing the use of Copilot in Bing and Image Creator from Bing are set forth in the **Code of Conduct** sections of Copilot AI Terms of Use and Image Creator Terms**.** A number of the provisions of the Code of Conduct prevent the creation of materials that could negatively impact human dignity. According to the code, users cannot: engage in activity that is harmful to themselves, the Online Services, or others; engage in activity that violates the privacy of others; engage in activity that is fraudulent, false, or misleading; infringe on the rights of others; use the service to create or share inappropriate content or material; or to do anything illegal. Users who violate the Code are subject to limitation from the service.

Terms and Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft

Services Agreement Code of Conduct broadly prohibits using Microsoft services to generate or share inappropriate content or material. In addition to the Microsoft Services Agreement, users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot in Bing's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with inappropriate content, including nudity, pornography, or violence).

Microsoft Advertising Policies: Microsoft Advertising, which powers ads on Bing, has clear and regularly enforced content policies and practices that prevent advertisements negatively impacting human dignity. The Microsoft Advertising Policies set out the requirements for ad content, including criteria upon which ad content will be removed. Microsoft requires its advertisers and partners to comply with its policies throughout their use of the Microsoft services. Microsoft Advertising also has a set of Relevance and Quality Policies to manage the relevance and quality of the advertisements that it serves through its advertising network.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and uses the authority as a part of QC score, which is one of Bing's main ranking parameters. Bing rigorously measures and monitors metrics to identify implementation gaps and inform product strategy.

Additional Algorithmic Interventions: Although Bing Search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or "defensive search interventions," to counteract these threats, in accordance with its search principles.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing's content removal practices for the search index are largely reactive to reports and involve human review. Bing's support teams are provided with training and oversight in order to ensure removal practices align with Bing's principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing's generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing’s product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing’s operations.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled “Feedback.” Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing’s internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing.

Honoring the Right to be Forgotten: Bing’s established “Right to Be Forgotten” policies and procedures help reduce risks of negative effects to human dignity by providing reporting tools for users to submit requests that certain online content associated with them be removed from Bing. Consistent with the EU Right to be Forgotten, EU users may submit requests to remove search results for queries that include the person's name where they can demonstrate the results are inadequate, inaccurate, no longer relevant, or excessive using Bing’s Request Form to Block Search Results in Europe. Bing has dedicated teams, training materials, and escalation paths for “Right to Be Forgotten” requests and removes indexed websites (and, as relevant, search suggestions) where the requestor or data subject’s privacy interest is not outweighed by the public interest. Bing also engages with local courts and Data Protection Authorities on data subject escalations of Right to Be Forgotten decisions and responds to feedback as appropriate.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of “classifiers” and “metaprompting” to help reduce the risk of harms and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers “low authority” content as part of an answer and to help address impermissible content and behaviors.

Safety Features: Bing also empowers users to control what they see within search results with features like **SafeSearch**. SafeSearch leverages advanced algorithms and machine-learning technologies to scrutinize and categorize content, ensuring that pornography and adult content are effectively filtered from search outcomes. SafeSearch is designed with user empowerment in mind, offering three adjustable settings: "Strict," which is the most protective filter, removing explicit text, images, and videos from results; "Moderate," the default setting, which filters explicit images and videos while allowing text that might not be suitable for certain ages; and "Off," which provides unrestricted access to search results. For added protection, especially in environments like schools or households with minors, SafeSearch settings can be locked, preventing unauthorized changes, and ensuring consistent enforcement of content filters. This flexibility allows individuals and organizations to customize their browsing experience in alignment with personal preferences and values or institutional policies. SafeSearch's underlying technology is continuously refined, employing the latest in machine-learning and content analysis to enhance its accuracy and effectiveness. Furthermore, Bing provides mechanisms for users to report inappropriate suggestions, enabling the community to contribute to the service's safety.

Product Improvement

Additional Algorithmic Interventions: Bing has a team accountable for implementing algorithmic interventions and metrics monitoring and remedying, via algorithmic interventions, high impact issues in search results, such as misinformation, hateful speech, and other problematic content that could negatively impact human dignity. Microsoft proactively scans images generated by Image Creator from Bing to detect any potentially harmful content in order to prevent the dissemination of harmful material. Bing regularly reviews the efficacy of its interventions to ensure that they are performing as expected and not inadvertently introducing additional bias or other harm. Bing's mitigations are designed to work across languages and markets where Bing is offered. While the robustness of the mitigations might vary depending on the traffic volume, Bing works to ensure the effectiveness of the mitigations by monitoring the metrics.

Monitoring Trends and Threats: Bing regularly monitors trends and emerging threats in order to evolve mitigation practices to meet the rapidly adapting tactics to circumvent Bing safeguards and access potentially harmful content within this risk category. Bing has adapted its service over the course of the assessment period to address these risks, including:

- Microsoft proactively tracks tactics and methods that users employ to jailbreak Copilot in Bing and Image Creator from Bing to produce content with a negative impact on Human Dignity and takes daily action to address any gaps in safeguards.
- Bing invests in classifiers, including adult classifiers and hate classifiers, to help identify queries that might lead to content that poses negative impact on human dignity.
- Microsoft regularly reviews the prevalence of issues within Image Creator from Bing to evaluate the landscape of potential harms to human dignity and adjust the safeguards accordingly. Through this process, Microsoft enhances the mitigations to reduce the occurrence of harmful outputs on human dignity.

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of Human Dignity on Bing's services.

External Collaboration: Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to help ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform the processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Product Transparency

Content Provenance and Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to help ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Limitations on Copilot Responses: Flags on Copilot in Bing by the classifiers have led to mitigations such as not returning generated content to the users, diverting the user to a different topic, or redirecting users to Search. For more information on Product Transparency, please see the Appendix II section, as the same mitigations apply to Human Dignity.

Public Security

Risks related to Public Security include risks related to terrorism and violent extremisms as well as coordination of harm and misinformation related to crisis events. Bing applies its comprehensive, multi-layered safety mitigations to content that poses risks to public safety. Bing's ranking algorithms and generative AI safety mitigation systems are designed to protect users from harmful web content and prevent harmful or misleading generative AI outputs. Bing's threat monitoring and incident response processes are designed to continuously monitor evolving trends and place rapid interventions where needed. Additionally, Bing's partnerships with organizations like the Institute for Strategic Dialogue (ISD) and its leadership role in the Global Internet Forum to Counter Terrorism (GIFCT) underscore its dedication to combatting terrorist, violent, and extremist content. Through these initiatives, Microsoft aims to contribute to the broader fight against online extremism and misinformation, aligning with Bing's commitment under the COPD to adapt to the challenges posed by generative AI. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Public Security, the Bing Risk Assessment team has assessed the Residual Risk related to Public Security on the service to be Moderate.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Public Security	Highly Likely	Critical	Critical	Managed	Moderate

Risk Definition

Risk Area	Public Security
Risk Definition	Risk that content or activities with actual or foreseeable negative effects on public security occur on the service.

Theoretical Risk Manifestations absent Mitigations	• Risk that Search results include terrorist recruitment content • Risk that Search results rank content promoting protests or mass activities of groups seeking to cause harm to others • Risk that activities on Shopping are used to collect funds that are used to finance terrorist organizations • Risk that Copilot in Bing responses include or link to information from low authority sources when providing responses to inquiries about crisis events • Risk that Image Creator from Bing is used to generate images that contribute to public security risks • Risk that bad actors bypass Bing's safety systems to increase the visibility of terrorist content • Risk that Copilot in Bing provides information on terrorism that enables terrorist activities • Risk that Image Creator from Bing is used to create Terrorist propaganda or imagery

Risk Analysis

The **Inherent Probability** of systemic risks related to Public Security stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Highly Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a "Highly Likely" level of probability relative to other potential systemic risks on Bing.
- Due to the nature of Bing as a search engine, Bing's goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less authoritative, while working to help ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing does not include features for users to post or share content within the service or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the Inherent Probability and ways that users may be exposed to potentially harmful content vary between core search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** for content and activities that negatively impact Public Security on Bing is rated as "Critical," primarily due to the gravity of risks within this category, considering the potential for significant harm to security, wellbeing, societal, political, and economic systems at the individual, country, and regional levels. The inherent impact of many risks within this category is irremediable.

With the rating for Inherent Probability as "Highly Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with Public Security on the Bing service is "Critical."

Risk Mitigation

Bing has implemented a robust set of mitigations to address risks related to Public Security, including those described in Bing's approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and summarized below. The mitigations Bing has implemented relative to Public Security follow best practices with defined, documented, and managed processes. As such, the **Maturity of Mitigation** efforts Bing has applied to the Public Security risk area is assessed as "Managed," which brings the **Residual Risk** rating for down to "Moderate."

While not a proxy for Residual Risk, Bing has measured the DDR for Public Security on Bing Search as an average of .74% from April to June 2024. This means that Bing estimates .74% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Public Security.

Considering the Moderate Residual Risk, Bing continues to prioritize investment in developing and enhancing strategies to further mitigate and manage risks related to Public Security. The key mitigations currently implemented are described below.

Product Development

Dedicated Expert Teams and Industry Engagement: Bing is actively engaged in enhancing public security through the collaboration of its internal teams, such as the Microsoft Threat Assessment Center (MTAC), and external experts specializing in public safety and misinformation. MTAC is a group of experts who analyze and report on nation state threats, including cyberattacks and influence operations. The team is responsible for detecting, assessing, and disrupting digital threats to Microsoft, its customers, and democracies worldwide. MTAC brings together a team of experts fluent in more than a dozen languages able to analyze both cyber and influence threats arising from nation states and the most prolific threat actors. Detecting emerging actors and methods used in malign influence operations is highly resistant to automated processes, but Microsoft has developed technology to pair the right data with the right team of human analysts. With data and geopolitical expertise on both the target populations and the entities engaging in malign influence, the team can detect new actors and methods—assessing and in many cases attributing influence activity. This cooperative effort is pivotal in swiftly identifying and addressing potential threats, especially in preparation for critical global elections including those in the EU. Through these measures, Bing demonstrates its commitment to maintaining a secure service by mitigating risks that could compromise public security.

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing and Image Creator from Bing and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

New Feature Launch: Bing relies on Microsoft's extensive cross-company compliance infrastructure to proactively review new products and features to ensure they meet Microsoft's policy commitments in key areas such as privacy, accessibility, digital safety, and RAI.

Product Governance

Terms and Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft Services Agreement Code of Conduct broadly prohibits using Microsoft services to engage in activity that is harmful to others. In addition to the Microsoft Services Agreement, Users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with harmful content, including posting terrorist or violent extremist content).

Incident Response Governance: As a member of GIFCT, an initiative to bring together technology companies, governments, and experts, to share means of countering terrorist and violent extremist from exploiting digital platforms, Bing has access to the GIFCT's Incident Response processes, including ingesting hashes related to an event activated as Content Incidents or Content Incident Protocols. This allows Bing to quickly become aware of, assess, and address potential content circulating online resulting from a terrorist or violent extremist event. Bing also participates in the Advisory Network (a multistakeholder advisory group) of the Freedom Online Coalition, a coalition of 37 governments working to advance Internet freedom.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to help ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of QC score, which is one of Bing's main ranking parameters. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

Additional Algorithmic Interventions: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or "defensive search interventions," to counteract these threats, in accordance with its search principles.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing's content removal practices for the search index are largely reactive to reports and involve human review. Bing's support teams are provided with training and oversight in order to help ensure removal practices align with Bing's principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing's generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing's product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing's operations.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to help ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of harms and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

Dedicated Expert Teams: Microsoft takes public safety and security seriously with dedicated teams like the MTAC. MTAC is a group of experts fluent in multiple languages, analyzing nation-state threats, cyberattacks, and influencing operations to protect Microsoft, its customers, and democracies worldwide. With data and geopolitical expertise on both the target populations and the entities engaging in malign influence, the MTAC team can detect new actors and methods—assessing and in many cases attributing influence activity. Furthermore, Bing has a team which is accountable for search principles creation, evaluation, and operations that curate intelligence on misleading and harmful content, coordinating responses to critical issues across Bing and other Microsoft web services. They provide transparency on actions taken and facilitate the adoption of content usage and moderation principles. The Information Integrity team conducts research and produces reports on threats, particularly analyzing the intersection between cyber-attacks and information influence operations. Finally, Bing has a dedicated team that is accountable for implementing algorithmic interventions and metrics monitoring and dedicated to identifying high-risk areas where Bing's ranking and relevance algorithms deviate from its principles and goals, through methods such as red team testing, external threat intelligence, and social listening systems. Their focus is on implementing algorithmic adjustments to address these discrepancies. The team is working on undertaking efforts to deploy interventions across mostly key languages and regions Bing operates in, and they regularly assess their mitigations using objective metrics to maintain the balance of not overblocking while providing high authority and quality search results.

Product Improvement

Industry Engagement: Bing works closely with the ISD to understand risks related to terrorist, violent, and extremist content (TVEC) in search, as well as to surface "counter narratives" via ad grants that help deter users who indicate an intent to learn more about extremist organizations.

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of negative effects to Public Security on Bing's services.

External Collaboration: Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to help ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform the processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or other special information "hubs" related to public health issues.

Content Provenance and Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to help ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Mental and Physical Wellbeing

Risks related to Mental and Physical Wellbeing include exposure to content pertaining to gender-based violence, harmful behavioral suggestions, and content that encourages self-harm like suicide. Bing applies its comprehensive safety mitigations to the wide range of web content that poses risks of negatively impacting users' mental and physical wellbeing. By prioritizing high authority content, Bing's ranking algorithms are designed to protect users from harmful web content, such as content that promotes suicide or self-harm, or content that advocates gender-based violence. Generative AI features deploy additional mitigations to prevent harmful AI-generated outputs. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Mental and Physical Wellbeing, the Bing Risk Assessment team has assessed the Residual Risk related to Mental and Physical Wellbeing on the service to be Moderate.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Mental and Physical Wellbeing	Likely	Critical	High	Defined	Moderate

Risk Definition

Risk Area	Mental and Physical Well-Being
Risk Definition	Risk that content or activities with actual or foreseeable serious negative consequences to a person's physical and mental well-being or in relation to gender-based violence occur on the service.
Theoretical Risk Manifestations absent Mitigations	• Risk that Copilot in Bing responses include content that could be distressing, upsetting, or otherwise harmful to the mental health of user • Risk that Image Creator from Bing is used to create content promoting gender-based violence or threatening violence • Risk that Copilot in Bing discusses or supports self-harm, when responding to users • Risk that Travel ideas, itineraries, and/ or search results suggest dangerous routes, destinations, or activities that place individuals in physical danger or otherwise undermine their physical well-being • Risk that Search is used to collect information about individuals to facilitate stalking, harassment, or other harm • Risk that Image Creator from Bing is used to create material encouraging or depicting self-harm or suicide • Risk that Search results or Autosuggest searches includes content that promotes or suggests materials related to suicide or self-harm • Risk that Advertisements appear on Bing that promote addictive or harmful products or illegal substances

	• Risk that Search or Image Creator from Bing results promote unrealistic or unhealthy body standards that could negatively impact mental health

Risk Analysis

The **Inherent Probability** of systemic risks related to Mental and Physical Wellbeing stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a "Likely" level of probability relative to other potential systemic risks on Bing.
- User intent is generally required for users to access potentially harmful content on Bing. Due to the nature of Bing as a search engine, Bing's goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less credible, while working to help ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** for content and activities that negatively impact Mental and Physical Wellbeing is rated as "Critical" primarily due to the gravity of risks within this category, considering the potential for significant harm to wellbeing, economic, and societal systems at the individual and potentially broader level. The impact of some risks within this category, such as violence against women, are irremediable. Other risks, such as behavioral addictions, may be considered remediable but still have the potential to cause significant damage to an individual's mental health or threaten their physical wellbeing.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with negative impacts to Mental and Physical Wellbeing on the Bing service is "High."

Risk Mitigation

Bing has implemented a set of mitigations to address risks related to Mental and Physical Wellbeing, including those described in Bing's approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and summarized below. The mitigations Bing has implemented relative to Mental and Physical Wellbeing follow best practices with defined and documented processes. As such, the **maturity of mitigation** efforts Bing has applied to the Mental and Physical Wellbeing risk area is assessed as "Defined," which brings the **Residual Risk** rating for down to "Moderate."

While not a proxy for Residual Risk, Bing has measured the DDR for Mental and Physical Wellbeing on Bing Search as an average of .71% from April to June 2024. This means that Bing estimates .71% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Mental and Physical Wellbeing.

Considering the Moderate Residual Risk, Bing continues to prioritize investment in developing and enhancing strategies to further mitigate and manage risks related to Mental and Physical Wellbeing. The key mitigations currently implemented are described below.

Product Development

Ranking Algorithms: Bing is committed to mitigate risks of mental and physical well-being such as gender-based violence, harmful behavioral suggestions, and content that encourages self-harm like suicide by continuously investing in ranking algorithms to prioritize high authority content in top search results. While protecting users against being exposed to such content, Bing also works to ensure users' rights to freedom of expression/information are respected by relying on ranking principles and algorithms rather than complete removal from index. This can potentially result in risks associated with mental and physical wellbeing to remain. Bing actively works on further improving mitigation techniques to combat these risks.

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing and Image Creator from Bing and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

Product Governance

Terms and Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft Services Agreement Code of Conduct broadly prohibits using Microsoft services to engage in activity that is harmful to others. In addition to the Microsoft Services Agreement, Users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with harmful content, including stalking, harassing, bullying, threatening, or advocating violence).

Microsoft Advertising Policies: Microsoft Advertising, which powers advertisements that appear on Bing, has clear and regularly enforced content policies that prevent advertising of harmful materials**.** Ads loaded into the Microsoft Advertising system are 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 advertisements flagged to its customer support team and removes advertisements that violate its policies. Microsoft Advertising's policies prohibit political advertising. Advertisers also retain ownership and responsibility for their ad content, but the advertiser must agree to Microsoft's terms when signing up for a Microsoft Advertising account. Microsoft Advertising monitors the service and removes ads and advertisers that violate Microsoft's agreement and policies. Microsoft Advertising employs dedicated operational support and engineering resources to enforce these policies, combining various enforcement methods to prevent or take down advertisements that violate its policies. Finally, Bing's advertising partner, Microsoft Advertising, likewise prohibits the display of ads in categories that may negatively impact physical or mental well-being. For example, among other types of restricted content, Microsoft Advertising has restrictions on advertisements: 1) advocating, glorifying, promoting, or facilitating any type of exploitation or human trafficking; 2) advertising services related to international matchmaking for marriage; 3) for spy-cams or surveillance equipment; 4) for weapons and firearms; 5) for people finder websites; and 6) for Adult websites/content. Microsoft uses both manual and systemic methods to review ads for compliance with their policies and reserves the right to remove those that are non-compliant.

Transparency to Users: Bing provides users more detailed information on the main parameters it uses for ranking in the How Bing Ranks Your Content section of the Webmaster Guidelines, specifically calling out that in measuring the "quality" of a website, "pages that call for violence, name-calling, offensive statements, or use derogatory language to make a point are generally considered low quality. Bing works to ensure these resources are available in relevant languages and markets across the EU.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to help ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of QC score, which is one of Bing's main ranking parameters. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

Additional Algorithmic Interventions: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or "defensive search interventions," to counteract these threats, in accordance with its search principles.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing's content removal practices for the search index are largely reactive to reports and involve human review. Bing's support teams are provided with training and oversight in order to ensure removal practices align with Bing's principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing's generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing's product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing's operations.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing.

Honoring the Right to be Forgotten: Bing's established "Right to Be Forgotten" policies and procedures help reduce risks of negative effects to mental and physical well-being by providing reporting tools for users to submit requests that certain online content associated with their name be removed from Bing. Consistent with the EU Right to be Forgotten, EU users may submit requests to remove search results for queries that include the person's name where they can demonstrate the results are inadequate, inaccurate, no longer relevant, or excessive using Bing's Request Form to Block Search Results in Europe. Bing has dedicated teams, training materials, and escalation paths for "Right to Be Forgotten" request and removes indexed websites (and, as relevant, search suggestions) where the requestor or data subject's privacy interest is not outweighed by the public interest. Bing also engages with local courts and Data Protection Authorities on data subject escalations of Right to Be Forgotten decisions and responds to feedback as appropriate.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of harms and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

Dedicated Expert Teams: Bing has a team that is accountable for implementing algorithmic interventions and metrics monitoring and dedicated to identifying and remedying, via targeted algorithmic interventions, high impact issues in search results, such as misinformation, public health concerns, hateful speech, self-harm materials, and other problematic content that could negatively impact these interests. Bing maintains a set of metrics to track, monitor, and review the efficacy of its interventions on an ongoing basis to help ensure that they are performing as expected and not inadvertently introducing additional bias or other harms.

Redirecting Search Queries: For Copilot in Bing, queries related to self-harm or suicide are considered high-risk, and Copilot in Bing responds to these queries through "thoughtful disengagement" and redirects users to search to reach out to family members, friends, and mental health professionals. For search queries regarding eating disorders, under Bing's content policies this falls under a form of self-harm, as it can be considered a physical risk for users, thus defensive interventions for user queries are employed to direct users towards high authority content. For responses generated by Copilot in Bing, there is a high authority approach taken regarding queries relating to eating disorders, to explain the negative effects of certain associated behaviors.

Product Improvement

Industry Engagement: Furthermore, Microsoft has developed tools and multistakeholder partnerships to combat the rise of disinformation that can pose risks to public health or negatively impact physical or mental wellbeing. Partnerships with the **Global Democracy Index**, **NewsGuard**, and **Reporters Without Borders** aim to empower Bing product teams to take additional actions to promote more authoritative information. Microsoft is also a founding member of the **C2PA** alongside Adobe, Intel, TruePic, Twitter, the BBC, and other tech and media companies. Earlier this year, the coalition launched the first version of its open-source content provenance tool, which allows creators to claim authorship while empowering consumers to make informed decisions about what digital media to trust.

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of negative effects to Mental and Physical Wellbeing on Bing's services.

External Collaboration: Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to help ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform the processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or other special information "hubs" related to public health issues.

Content Provenance and Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to help ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Internal Subject Matter Experts: Bing partners with, subject matter expert teams tasked with assisting Microsoft services and products in developing an understanding of these types of risks. Furthermore, The Digital Safety Unit and the Office of RAI are two such teams that were established to support safety and responsible AI practices across Microsoft products and services, including Bing.

Directing to Authoritative Sources: Bing adds information to help ensure that users are not harmed by materials returned to search queries and will point to authoritative information, when search results lack high authority content, or redirect users when they search for specific queries. For example, when users search for information on suicide methods they are redirected to suicide prevention hotlines, which are available in 31 countries (including certain EU member states). This also applies to images or videos related to suicide, and search queries self-harm, self-harm images or videos. Queries related to self-harm or suicide are considered high-risk queries, and harmful content around these queries are considered low quality, and generally demoted but demoting user-generated content is tricky for image and videos due to quality control signals from social media but the PSA to redirect users to help lines are shown. Similarly, PSAs are also shown for user search queries to Domestic Violence, directing users to hotlines and report.

Rights and Protection of Minors

Risks related to Rights and Protection of Minors include the sharing of content pertaining to Child Sexual Abuse Material (CSAM) or Child Sexual Exploitation and Abuse Imagery (CSEAI), exposure of minors to harmful behavioral activities and content, or collection of minor's data without parental consent. Bing is committed to prioritizing the rights and safety of minors, aligning its product designs with the principles of the UN Convention on the Rights of the Child (UNCRC). Bing deploys proactive scanning to prevent CSEAI from entering into the Bing index. To protect young users from harmful web content, Bing offers a SafeSearch feature to enable filtering out adult and graphic or violent content. SafeSearch is available without authentication. Additionally, Microsoft Family Safety features allow parents to set their minor's search experiences to a SafeSearch setting.

Bing does not require user authentication for access to search. This is standard practice across search engines and is of the utmost importance given the role that search engines play in user free access to information. Allowing users to search without authentication is also an important privacy consideration for both adults and minors as authentication requires users to submit personal information. Bing search users who are not identified through authentication as minors are treated as adults.

Microsoft believes there is value in minors engaging with generative AI services. At the same time, Bing also understands that, without appropriate safeguards, minors may engage with these tools in harmful or undesirable ways. In the interest of rights and protection of minors, full generative AI services are available only to authenticated users over the age of 13 (or higher in jurisdictions with a higher age of parental consent). Any user signed in with a Microsoft account that identifies the user as under 13 years of age (or older in jurisdictions that require parental consent for older teens) cannot access these services in an authenticated state, even in the limited version available to unauthenticated users. Unauthenticated users are offered only a limited preview of the Copilot in Bing service that allows for a small number of conversations/turns per conversation. If SafeSearch is activated, even this limited preview is not available. Image Creator from Bing is disabled for all unauthenticated users.

Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Rights and Protection of Minors, the Bing Risk Assessment team has assessed the Residual Risk related to Rights and Protection of Minors on the service to be Moderate.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Rights and Protection of Minors	Likely	Critical	High	Defined	Moderate

Risk Definition

Risk Area	Rights and Protection of Minors
Risk Definition	Risk that content or activities with actual or foreseeable negative effects on the protection minors or respect for the rights of the child occur on the service.

Theoretical Risk Manifestations absent Mitigations	• Risk that users search for CSAM using Bing • Risk that autosuggestions on Search expose minors to harmful third-party content, such as hate speech, extremist content, violent or graphic material, adult content, information promoting or soliciting Illegal and regulated goods or self-harm • Risk that Copilot in Bing generates harmful content or content promoting self-harm to minor users • Risk that Image Creator from Bing generates harmful content inappropriate of/for minors • Risk that Bing’s data processes harm minors • Risk that Advertisements on Bing expose minors to harmful content, or includes content containing, child sexual exploitation and abuse, or CSAM • Risk that Advertisements on Bing target minors • Risk that Search results recommend risky behaviors among minors, such as dangerous challenges or trends from social media platforms • Risk that Bing’s safety systems unduly limit minors’ access to information • Risk that minors interacting with Copilot in Bing self-generate sexualized material • Risk that Copilot in Bing generates content relating to abuse and/or grooming a minor

Risk Analysis

The **Inherent Probability** of systemic risks related to the Rights and Protection of Minors stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a “Likely” level of probability relative to other potential systemic risks on Bing.
- User intent is generally required for users to access potentially harmful content on Bing. Due to the nature of Bing as a search engine, Bing’s goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less credible, while working to help ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.
- It is important to note that the Bing service is not known to be used by a large number of minors (less than 1% of EU Bing MAU are known to be users under 18), and minors under the legal age of

consent in their respective location are prevented from accessing the full slate of Bing generative AI features.

The **Inherent Severity** for content and activities that negatively impact the Rights and Protection of Minors on Bing is rated as "Critical" due to the gravity of risk in this category, the vulnerable nature of the population impacted, the irremediability of impact of certain risks, namely CSAM, and the potential for societal impact broader than the individual scale for risks impacting youth. Not all risks within this category are assessed at the highest level of severity, namely collection, processing, and use of minor's data without parental consent.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with negative impacts to the Rights and Protection of Minors on the Bing service is "High."

Risk Mitigation

Bing has implemented a set of mitigations to address risks related to the Rights and Protection of Minors, including those described in Bing's approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and summarized below. The mitigations Bing has implemented relative to the Rights and Protection of Minors follow best practices with defined and documented processes. As such, the **Maturity of Mitigation** efforts Bing has applied to the Rights and Protection of Minors risk area is assessed as "Defined," which brings the **Residual Risk** rating for down to "Moderate."

While not a proxy for Residual Risk, Bing has measured the DDR for the Rights and Protection of Minors on Bing Search as an average of .73% from April to June 2024. This means that Bing estimates .73% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to the Rights and Protection of Minors.

Considering the Moderate Residual Risk, Bing continues to prioritize investment in developing and enhancing strategies to further mitigate and manage risks related to the Rights and Protection of Minors. The key mitigations currently implemented are described below.

Product Development

Safety Features: Bing offers several features that help prevent child accounts from inadvertently accessing harmful content and put parents and other caregivers in control of the type of content their families encounter online. Parents can use the Microsoft Family Safety Center to block features across Microsoft for their children. The safety tools within the center allow for monitoring of search, screen time, and to set certain filters for their children, including in SafeSearch which filters explicit images and text. The **SafeSearch** feature allows Bing users (and for those using Family Safety features, other users in their "family" account) to control what type of adult content may appear in search results. SafeSearch is set by default to "moderate" in most markets, which restricts the display of adult imagery. "Strict" mode prevents the display of both explicit text and images. "Off" mode allows for full display of content. Additionally, parents can use family safety settings in SafeSearch to disable Bing's generative AI features on a device or a network. Family Safety settings also allow parents to track their family's search history and monitor screen time. Additionally, network administrators (including schools and parents) can fully disable chat via instructions provided here. Child users whose parents have access to their search history are aware when these controls are in place. Microsoft Family Safety services are documented here: View device and app use with activity reporting - Microsoft Support.

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing and Image Creator from Bing and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

User Feedback: In recognition of the critical importance of understanding and addressing the needs and safety of minors online, Microsoft has taken proactive steps to engage directly with youth and gather their insights. For example, Microsoft, with support from European SchoolNet, hosted a session with Better Internet for Kids Youth Ambassadors to discuss their experiences with AI-powered search. This initiative, alongside research partnerships with organizations like 4H in the US, underscores Bing's commitment to better protecting and serving young users, particularly those from at-risk communities.

Honoring the Right to be Forgotten: Bing's established "Right to Be Forgotten" policies and procedures help reduce risks to the rights of the child or minors by providing reporting tools for users to submit requests that certain online content associated with their name be removed from Bing. Consistent with the EU Right to be Forgotten, EU users (or their representative) may submit requests to remove search results for queries that include the person's name where they can demonstrate the results are inadequate, inaccurate, no longer relevant, or excessive using Bing's Request Form to Block Search Results in Europe. Bing has dedicated teams, training materials, and escalation paths for "Right to Be Forgotten" request and removes indexed websites (and, as relevant, search suggestions) where the requestor or data subject's privacy interest is not outweighed by the public interest. Bing also engages with local courts and Data Protection Authorities on data subject escalations of Right to Be Forgotten decisions and responds to feedback as appropriate.

Algorithmic Interventions: Like other risk areas, Bing utilizes targeted algorithmic interventions to remedy high impact issues in search results that could negatively impact minors and reviews the efficacy of the interventions.

Microsoft Privacy Dashboard: All authenticated users, including teen accounts, have data subject rights over data collected by Microsoft available on the Microsoft Privacy Dashboard. For conversational data in

Bing, users can view individual initial conversation triggers and can export and delete product and service usage data to delete stored conversation history. Where Bing asks for consent for data collection and use, it considers whether consents are written in language a child can understand, and defaults settings to a high level of privacy protection. In appropriate cases, Bing has fully disabled consents for child users (e.g., optional data collection through Bing's EU privacy banner).

Data Protection Impact Assessment (DPIA): Bing completes full DPIAs for any new products or features; the DPIA requires specific investigation of possible harms to data subjects based on processing of their personal data and documentation that Bing has identified an appropriate legal basis for processing that data under General Data Protection Regulation (GDPR) as well as to explain how Bing has mitigated possible privacy risks to users. The DPIA process also requires a review as to whether the service has appropriately addressed possible harms to minor users. Although, to Bing's knowledge, there are relatively few minors on the Bing service, Bing still works to comply with applicable Microsoft policies and data use requirements as to minor users, and to consider the best interests of the child in product development. For known child users, Bing sets default controls to the highest level of privacy protection, such as turning off optional data collection by default in Bing's cookie banners. Additionally, users are not permitted to search for adult content using the visual search feature, which helps limit the possibility that users could find CSEAI on the service.

New Feature Launch: Regarding improvements to products Bing relies on Microsoft's extensive cross-company compliance infrastructure to proactively review new products and features to ensure they meet Microsoft's policy commitments in key areas such as privacy, accessibility, digital safety, and RAI. For example, privacy reviews are an essential part of any new feature review process; implementing recommended privacy mitigations following such reviews is a requirement for launch. Microsoft has a robust privacy and security infrastructure, consisting of privacy managers who are trained in Microsoft privacy standards and relevant laws, and have access to centralized privacy specialist teams and legal support for complex or novel issues. Regarding minors there are specific standards that must be adhered to. For example, privacy reviews ensure that minor users are defaulted to the highest level of privacy protection and that disclosures and consent experiences are written in child-friendly language.

Generative AI Feature Testing: Copilot in Bing has been designed to avoid production of potentially offensive, harmful, or illegal materials that could negatively affect the wellbeing or development of teen users. Before launching, Copilot in Bing underwent extensive testing and reviews to ensure compliance with Microsoft's RAI principles, market-specific legal requirements, and Bing's trustworthy search principles.

Product Governance

Microsoft Privacy Statement: Users can learn more about Microsoft's collection and use of personal data in the Microsoft Privacy Statement. The Microsoft Privacy Statement outlines what personal data Microsoft collects, how it is used, purposes for which it is used, and how to access and control personal data. This includes how these components apply to specific Microsoft products like search and browse, cookies, Microsoft Account information, and minors' data. Additional topics about personal data are covered such as security, storage, and retention, and how to contact Microsoft. Additionally, Microsoft offers a child-friendly version of its privacy statement to better inform younger users of data practices. Microsoft has also developed a bespoke privacy page for young people that talks to Microsoft's data collection practices in age-appropriate language.

Terms and Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft Services Agreement Code of Conduct broadly prohibits using Microsoft services to engage in activity that exploits, harms, or threatens to harm minors. In addition to the Microsoft Services Agreement, Users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with content that is harmful to minors, including CSAM and non-age-appropriate content).

Microsoft Advertising Policies: Bing has implemented a comprehensive suite of measures aimed explicitly at safeguarding the rights, privacy, safety, and security of minors across its service. These measures include enforcing strict advertising policies through Microsoft Advertising, setting age restrictions under the MSA, developing age-appropriate privacy communications, and ensuring Bing features are accessible and understandable to teen users. Additionally, Bing's commitment to privacy protection for minors is evident through practices such as prohibiting targeted advertising to users under 18, minimizing data use, and engaging with external stakeholders on children's safety and development. Through these efforts, Bing demonstrates its dedication to creating a secure and supportive online environment for young users, while continuously seeking to enhance its protective mechanisms in alignment with evolving standards and stakeholder feedback, including Microsoft policy teams, regulators, and civil society, to continue to iterate and improve on trust and safety in the Bing service.

Microsoft Advertising Policies for Minors: Microsoft does not deliver personalized advertising to minors whose birthdate in their Microsoft account identifies them as under 18, in accordance with the **Microsoft's Privacy Statement: Children and advertising**. Across Bing features, authenticated users under 18 are not subject to targeted advertising. This important protection for youth will be carried forward into the new Bing features as well. Per the Microsoft Advertising Remarketing policy, advertisers must also not (i) "create a remarketing list, retarget or otherwise profile" any individuals under the age of 18; (ii) target individuals under the minimum age required for the product advertised; and (iii) implement remarketing on sites or applications directed to minors under the age of 18 or other applicable age limitations in an applicable market. Similarly, Microsoft has clear policies in place to prevent showing ads that are behaviorally targeted or contain adult materials such as alcohol or gambling to known child users. Microsoft uses both manual and systematic methods to review ads for compliance with their policies and reserves the right to remove those that are non-compliant.

External Consultations: Microsoft routinely engages with external experts on child online safety, through bilateral conversations and in forums such as the Internet Governance Forum (IGF). For instance, at the IGF in October 2023, Microsoft representatives participated actively in multiple sessions on child safety topics, with a particular focus on children's rights online. Additionally, Microsoft is a founding and active member of the Tech Coalition, focused on facilitating industry cooperation to address online child sexual exploitation and abuse, thus demonstrating a strong commitment to collaborative efforts in combating evolving CSEA challenges, where Microsoft routinely receive insights and information from other industry participants, as well as briefings from researchers. Additionally, Microsoft participated in the Civil Society Dialogue FY24 and explored two key topics; state privacy legislation, and minor's data and AI, with civil society partners. Microsoft presented some draft guiding principles and foundational capabilities for developing AI products and services for young people. Finally, feedback from external engagements, such as the Family Online Safety Institute's positive review of Copilot for Teens, has provided valuable insights into enhancing the user experience for young audiences. These include suggestions for improving geolocation accuracy and the readability of responses, furthering Bing’s commitment to transparency and reliability in its digital offerings.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing’s risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to help ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of Quality and Credibility score, which is one of Bing’s main ranking parameters. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

Additional Algorithmic Interventions: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or “defensive search interventions,” to counteract these threats, in accordance with its search principles.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing’s content removal practices for the search index are largely reactive to reports and involve human review. Bing’s support teams are provided with training and oversight in order to help ensure removal practices align with Bing’s principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled “Feedback.” Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing’s internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of “classifiers” and “metaprompting” to help reduce the risk of harms and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the A model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

Generative AI Features: Copilot in Bing features have been designed to avoid production of potentially offensive, harmful, or illegal materials that could negatively affect the wellbeing or development of teen users. Bing works to ensure that its generative features appropriately balance safety considerations and the risk of overblocking outputs so that users are able to access the information they seek. Copilot in Bing also provides several touchpoints for meaningful AI disclosures, where users are notified that they are interacting with an AI system and are presented with opportunities to learn more about these features and generative AI, such as through in-product disclaimers, Copilot in Bing: Our Approach to Responsible AI, educational FAQs, and blog posts. Empowering users with this knowledge can help them avoid over-relying on AI and learn about the system's strengths and limitations. Bing's in-product disclosures are written at a U.S. 5th grade level (where children are typically 10-11 years old) to maintain understandability by teen users.

User Restrictions: Copilot in Bing and Image Creator from Bing are available to authenticated users over the age of 13 (or higher in jurisdictions with a higher age of parental consent). Any user signed-in with a Microsoft account that identifies the user as under 13 years of age (or older, in jurisdictions that require parental consent for older teens) cannot access these services in an authenticated state, even in the limited version available to unauthenticated users.

Accessibility: Bing features are working towards meeting Microsoft's standards for accessibility, which will benefit any teen users who require additional assistance.

Policy Violating Content Monitoring: Bing proactively uses hash-matching technologies (including PhotoDNA and MD5) to detect matches to known CSEAI, to avoid it from appearing in the search index, reactive reporting, and removal, and via threat intelligence provided by third-party expert partners. Furthermore, Bing relies on the same robust reporting and reactive infrastructure, as previously mentioned to action illegal or policy violating content.

Internal Trainings: Bing's Content Moderation Ops team carries out training to ensure alignment on policy, labelling, and enforcement. For high-consequence harms, like child sexual exploitation and abuse, specialized teams receive additional focused training and wellness resources.

Product Improvement

Additional Safeguards and Enhancements for Minors: While Bing has implemented measures to protect minors from certain online harms, Bing is aware that there are areas where enhancements could further bolster these protections. Given that the vast majority of Bing users are unauthenticated and thus default to being treated as adults, minors may inadvertently be exposed to content and recommendations unsuitable for their age, for example such as tattoo shops and bars that are surfaced by the Maps Local Guide feature. To address this gap, additional safeguards and enhancements tailored specifically for authenticated minor users should be implemented to maintain a safer and more appropriate online environment across Bing services.

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of negative impacts to the Rights and Protection of Minors on Bing's services.

External Collaboration: Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to help ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform the processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. These experiences can help direct younger users to high authority information faster and aid in learning and research.

Content Provenance & Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Digital Literacy: Furthermore, Microsoft is dedicated to improving digital literacy online to help minimize the likelihood that users are harmed by misinformation or other problematic content appearing in Microsoft products, including Bing search results or conversational features.

Transparency Reports: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in EU member state languages. Microsoft offers a child-friendly version of its privacy statement to better inform younger users of data practices. Where content is removed from search, Bing is transparent by notifying users through a notice on the search engine results page and publishing regular transparency reports (available on Reports Hub | Microsoft CSR). Microsoft's Reports Hub is a publicly available source where users can access various transparency reports in areas where Bing reports on various practices. These key areas include: RAI, content removal requests (such as copyright or digital safety), privacy (such as "Right to be Forgotten") and security (such as Government Requests.) The Reports Hub contains additional transparency reports such as jurisdictional, community and privacy and security transparency reports that users can easily access. As a result of the conduct of the Systemic Risk Assessment, each year Bing identifies specific areas for focused enhancements in the coming year

In-Product Indicators: For users search queries for Child Sexual Abuse Material Bing displays PSAs, similar to those regarding negative searches regarding suicide, stating that Child sexual abuse material is illegal, but also includes links to report these illegal materials or receive anonymous support, aimed at those who are victims of CSAM. For image and video queries relating to CSAM, no results are shown, and instead shows unrelated images from popular web content, and for Copilot in Bing, the response to CSAM related queries is disengagement stating that the request is inappropriate or harmful.

AI Disclosures: Copilot in Bing and Image Creator from Bing contain in-product disclosures to make it clear to users that they are engaging with an AI system and remind that AI can make mistakes. These are written to be consumable across ages. In addition, product FAQs, help pages, and other public facing information sources help educate users on the nature of AI-driven search experiences and the uses, safeguards, and limitations of this emerging technology, regularly reminding users of the potential for mistakes and risks of over-reliance, which is critical for users of all ages.

Additional AI Protections: Bing has engaged in extensive RAI reviews regarding generative AI features in order to minimize the likelihood of harm. Bing prevents known users under the age of consent from accessing the full slate of conversational AI features.

Areas of Low Residual Risk

The remaining Risk Areas were assessed at a Residual Risk rating of Low, demonstrating alignment between Bing's investment in mitigations and risks on the service.

Consumer Protection and Fraud

Risks related to Consumer Protection and Fraud include exposure to content pertaining to scams, spam, false representation and information, fraudulent businesses, bots, deceptive commercial practices. Bing's ranking algorithms are designed to prevent users from being exposed to low quality content, including websites engaging in scams or fraudulent activities. Bing's spam policies and detection protect users from web spam that manipulates search algorithms and does not add content value. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Consumer Protection and Fraud, the Bing Risk Assessment team has assessed the Residual Risk related to Consumer Protection and Fraud on the service to be Low.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Consumer Protection and Fraud	Highly Likely	High	High	Managed	Low

Risk Definition

Risk Area	Consumer Protection and Fraud
Risk Definition	Risk that content or activities with actual or foreseeable negative impact to a high-level of consumer protection occur on the service.

Theoretical Risk Manifestations absent Mitigations	• Risk that advertisements link to malware • Risk that Search or Maps results include content that misleads consumers about the legitimacy of businesses, products, or services • Risk that Shopping listings include fraudulent, deceptive, or unsafe products • Risk that Copilot in Bing results reference or link to scam businesses or fraudulent websites • Risk that Image Creator from Bing is misused to generate highly realistic images of non-existent individuals for use in phishing schemes, fake endorsements, or efforts to create synthetic identities that could be used to defraud businesses and consumers • Risk that Copilot in Bing is misused to allow for the generation of content that can be used to scam, defraud, or spam individuals with more efficiency

Risk Analysis

The **Inherent Probability** of systemic risks related to Consumer Protection and Fraud stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Highly Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a “Highly Likely” level of probability relative to other potential systemic risks on Bing.
- Due to the nature of Bing as a search engine, Bing’s goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less credible, while working to ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results, and spam content is removed upon spam detection. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** of impact for fraudulent content or activities on the Bing service is rated as "High" due to the gravity of risks within this category and the potential for significant economic impact at the individual-level. While economic, and even security, societal, and wellbeing, impact at the individual-level can be severe, the risk is generally limited to the individual-level and does not necessarily scale to the country, regional, or global impact scale. Not all risks within this category would be considered "High," as the impact of spam and deceptive advertisements are less likely to cause significant harm, compared to malware links or deceptive Job postings.

With the rating for Inherent Probability as "Highly Likely" and Inherent Severity as "High," the **Inherent Risk** associated with Consumer Protection and Fraud on the Bing service is "High."

Risk Mitigation

Bing has implemented a robust set of mitigations to address risks related to Consumer Protection and Fraud, including those described in Bing's approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and those summarized below. The mitigations Bing has implemented relative to Consumer Protection and Fraud follow best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts Bing has applied to the Consumer Protection and Fraud risk area is assessed as "Managed," which brings the **Residual Risk** rating for down to "Low."

While not a proxy for Residual Risk, Bing has measured the DDR for Consumer Protection and Fraud on Bing Search as an average of .82% from April to June 2024. This means that Bing estimates .82% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Consumer Protection and Fraud.

Nevertheless, Bing continues to invest in developing and enhancing strategies to further mitigate and manage risks related to Consumer Protection and Fraud. The key mitigations currently implemented are described below.

Product Development

Spam & Abuse Monitoring: Spam tactics are often used by bad actors to manipulate the ranking and show up in search results. Bing invests significant time and resources to ensure that users are provided with high quality content to avoid the possibly of its page crawlers and algorithms inadvertently returning results that could negatively impact consumers, such as spam or other fraudulent websites. Bing continuously monitors manipulation trends in high-risk areas and deploys mitigations to ensure that quality results are returned to users.

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing, and Image Creator from Bing, and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

Feature Evaluation: Furthermore, Bing relies on its extensive compliance infrastructure, defensive search, and quality assurance teams to ensure that new features and metrics are evaluated routinely and are functioning to design.

Product Governance

Terms & Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft Services Agreement Code of Conduct broadly prohibits using Microsoft services for misleading or deceptive purposes. In addition to the Microsoft Services Agreement, Users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with deceptive content, including malware, spam, and phishing material).

Microsoft Advertising Policies: Microsoft Advertising, which powers ads on Bing, has clear and regularly enforced content policies and practices that prevent advertisements that may negatively affect consumer protection. The Microsoft Advertising Policies set out the requirements for ad content, including criteria upon which ad content will be removed. Microsoft requires its advertisers and partners to comply with its policies throughout their use of the Microsoft services. 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. For consumer protection, Bing's Webmaster Guidelines, anti-spam/manipulation policies, and sensitive information policies are in place to address the risks of materials that might impact consumer data protection and could be used for scams or fraud, such as pages containing account information or credit card numbers. In addition to detection methods for spam and malware content, Bing's user report mechanism works to address users' concerns with specific websites appearing in Bing search results.

External Consultation: Moreover, Bing regularly meets with regulators around the world, including the European Commission and EU member state Digital Service Coordinators, to understand key concerns, share information, and incorporate feedback into product design and safety systems as appropriate.

- For example, Bing responded to a consumer protection report released by a (non-EU) government agency concerning the ability of bad actors to use search engines to find websites offering stolen credit card details and other sensitive information that criminals can use to defraud people. As part of its response, Bing undertook measures to address issues raised in the reported, including removal of certain identified websites for policy violations and adjustments to related search suggestions and committed to continue undertaking efforts to address risks that search results can lead users to sites involved in illicit activities, such as credit card fraud.
- Bing also relies on internal and external authoritative sources for improving ranking capabilities to ensure that potentially harmful materials, spam, and low quality sites are lowered in ranking in user search results.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of QC score, which is one of Bing's main ranking parameters. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

Additional Algorithmic Interventions: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or "defensive search interventions," to counteract these threats, in accordance with its search principles.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam, and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing's content removal practices for the search index are largely reactive to reports and involve human review. Bing's support teams are provided with training and oversight to ensure removal practices align with Bing's principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing. Bing specifically allows users to report malicious websites, such as malware, phishing, spam, or exploitative content removal practices and actions user reports consistent with Bing policy.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of "classifiers," and "metaprompting" to help reduce the risk of harm and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses, or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the A model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers “low authority” content as part of an answer and to help address impermissible content and behaviors.

Microsoft Advertising Detection: The Microsoft Advertising detection system is multi-tiered, with checks based on main line (“ML”) models, heuristic sweeps/rules, and manual reviews to prevent bad actors from coming onto the system. Overall, Bing has models and rules that apply at the advertiser-level as well as those that apply to a particular ad.

- Some of the models that apply at the advertiser-level include:
    - Payment fraud detection: This model triggers when the advertiser adds a payment instrument. This model is owned by the commerce risk team and Bing consume its decision.
    - Purchase risk model: This model triggers upon billing the advertiser to determine whether to try to make the charge or block the advertiser. This helps to bring up-to-date information to determine payment risk.
    - Behavioral risk model: This model looks at various features of an advertiser’s account or campaign(s) such as the ad text, keywords, budgets, account age, account structure, attributes of the payment instrument, individuals associated with the account, devices from which the advertiser is accessing the system etc. to determine the risk of an advertiser being fraudulent. This model runs when the advertiser makes any change to their ad campaigns (ads/keywords/budgets etc.).
    - LLM-based spoofing detection: One of the large categories of fraud is where advertisers pretend to be a legitimate site (e.g., trying to deceive users into believing that an ad leads to the official website of a bank). Bing has an LLM-based approach looking at advertiser demand to try and predict if advertiser is engaging in this behavior.

- Bing applies models to individual ads based on the corresponding Microsoft Advertising Policies applicable to the market where the ads serve. Such models include, among others: Clickbait, adult, dating, sensitive topics (hate speech, violence, gore), gambling.

- Based on the above models' output, Bing decides to either close the advertiser or send it to manual review for a decision. In addition to the models described above, Bing evaluates and uses other signals from Microsoft internal and external providers. For example, Microsoft Advertising uses a signal from the Bing team that gives an indication of whether a particular domain has spam. Bing uses signals from the Defender team which indicates that the user would be harmed by visiting this page.

    - Some of the signals from external providers include:
        - Signals from VirusTotal, an aggregator of malware risk signals from multiple providers.
        - Domain tools – which have indicators of domain risk.
        - Who is – which provides useful information such as age/expiry of domain, clicks on the domain on Bing and Google.

Real Estate Vertical Fraud Detection: In this assessment period, Bing Real Estate expanded its capabilities to automatically detect fraudulent content. With these expanded capabilities, Bing Real Estate can better detect new trends in fraudulent content. Bing continues to improve Real Estate's fraud detection capabilities as new methods for posting harmful content arise.

Webmaster Abuse Monitoring: Bing's Webmaster team invests significant resources in addressing attempts by webmasters to engage in abusive Search Engine Optimization (SEO) techniques and Bing's ranking systems are designed specifically to prioritize high authority content, rather than websites deploying spam or other manipulative tactics. Bing teams regularly monitor trends and emerging threats to evolve mitigation practices to meet the rapidly adapting fraud, mainly manifested by spammers attempting to manipulate search results and Copilot in Bing responses. Additionally, Bing's general abuse/spam policies, detailed in Bing's Webmaster Guidelines, prohibit certain practices intended to manipulate or deceive the Bing search algorithms.

Dedicated Expert Teams: Bing has a dedicated team that is accountable for implementing algorithmic interventions and metrics monitoring. These interventions target high impact issues in search results where users have a heightened likelihood of being harmed, such as scams/fraud, misinformation, and other problematic content that could negatively impact consumer protection. Bing regularly reviews the efficacy of its interventions to ensure that they are performing as expected and not inadvertently introducing additional bias or other harm. Bing's defensive interventions are measured across markets and languages.

Spam Detection: Bing leverages spam detection systems that include dozens of Machine-learning models and heuristics-based pipelines which detect Spam at a single webpage level, and across multiple webpages/sites. These Spam detection systems operate at an Internet-scale (tackling hundreds of billions of URLs) and remove spam sites manipulating search ranking and engaging in fraudulent activities.

Microsoft Advertising Policy Enforcement: Understanding the critical importance of maintaining a trustworthy ad ecosystem through Microsoft Advertising enforces stringent content policies that strictly prohibit misleading, deceptive, and fraudulent advertisement content. This is complemented by a set of **Relevance and Quality Policies** aimed at ensuring the served advertisements are not only legal but also relevant and of high quality, thus preventing advertisers from utilizing questionable or misleading tactics to lure users. Advertising employs a combination of human and automated review process using algorithms systems and a robust filtration system to detect and mitigate harmful online traffic such as fraud, phishing, malware, or hacked accounts and Bing users can report advertising for takedown through the **"feedback."** Detection on violative ad content covers ad text, images, extensions product feeds and components, landing pages, URLs, keywords, and targeting settings. Malware specifically is considered an egregious violation and will result in the suspension of an advertising account.

Microsoft Advertising Merchant Center: Products in Shopping, primarily leverage product advertisements from Microsoft Advertising. Microsoft reviews each listing added to the Microsoft Merchant Center product feed. Product listings that do not comply with Microsoft Advertising Policies are disapproved. Any merchant that violates Microsoft Advertising Policies will also be disapproved, in this case no products from the merchant will be visible. Bing's Shopping team conducts another layer of review on top of Microsoft Advertising when including products within Shopping search results.

Product Improvement

Microsoft Advertising Signals Improvement: Microsoft Advertising continually seeks to add new signals and information sources to improve its detection systems, evaluating the incremental performance of adding these signals, and sources to sweeps and models and adding those that it finds useful to the system.

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of Fraud and negative impacts to Consumer Protection on Bing's services.

External Collaboration: Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to help ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform the processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or other special information "hubs" related to public health issues.

Content Provenance & Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Public Service Announcement: In certain circumstances, Bing may add PSAs at the top of Bing search results to point users to high authority information on a particular topic or provide in-product warnings on particular URLs known or believed to contain harmful information (e.g., unaccredited online pharmacies and sites containing malware). PSAs appear as answer boxes at the top of a list of search results for certain categories of queries, providing information on potential risks associated with that query.

Protection of Personal Data

Risks related to the Protection of Personal Data include exposure to content pertaining to collection, processing, release of data, targeted ads, hacking, malware, phishing, data breaches, insufficient protection of data, and unauthorized disclosure or exposure of personal data.

Microsoft's longstanding belief that privacy is a fundamental human right has informed every stage of Microsoft's development and deployment of the Copilot in Bing experience. Microsoft's commitment to protecting the privacy of users, including by providing individuals with transparency and control over their data and integrating privacy by design through data minimization and purpose limitation, are foundational to Copilot in Bing. As Microsoft evolves its approach to providing the Copilot in Bing's generative AI experiences, it will continually explore how better to protect privacy. More information about how Microsoft protects its users' privacy is available in the Microsoft Privacy Statement.

Bing adheres to Microsoft privacy standards and relevant laws. Bing completes full DPIAs for any new products or features that involve materially different personal data processing. Bing also adheres to Microsoft-standard security practices that protect stored data from inappropriate access, via both technical and organizational means. Bing limits the retention of data to that necessary to provide the service. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to the Protection of Personal Data, the Bing Risk Assessment team has assessed the Residual Risk related to the Protection of Personal Data on the service to be Low.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Protection of Personal Data	Highly Likely	High	High	Optimized	Low

Risk Definition

Risk Area	Protection of Personal Data
Risk Definition	Risk that content or activities with actual or foreseeable negative effects on the protection of personal data occur on the service.
Theoretical Risk Manifestations absent Mitigations	• Risk that malicious actors gain unauthorized access to personal information via user prompts in Copilot in Bing • Risk that users are harmed by inappropriate processing of personal data, such as lack of transparency, control, access to remedies, or disclosures • Risk that Search results enable easy retrieval of information about an individual that falls within their private sphere and that the individual did not intend or consent to make available publicly, such as extremely sensitive personal information that could create risks of identity thefts such as credit card numbers, medical records, etc. • Risk that information requested to be removed under a "Right to be Forgotten" request is incorrectly processed or technical issues impact removal of content subject to a valid removal request • Risk that GenAI features inaccurately interpret source data to provide incorrect or misleading information about an individual • Risk that Search history data, if inadvertently disclosed, reveals an individual's beliefs, relationships, sexuality, medical issues, or other private information • Risk that "Right to Be Forgotten" requests from EU users are improperly actioned or technical issues prevent removal of URLs subject to a valid removal request

Risk Analysis

The **Inherent Probability** of systemic risks related to the Protection of Personal Data stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Highly Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a “Highly Likely” level of probability relative to other potential systemic risks on Bing.
- Due to the nature of Bing as a search engine, Bing’s goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less credible, while working to ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** for content and activities that negatively impact Protection of Personal Data is rated as “High” primarily due to the gravity of risks within this category, considering the potential for significant harm to economic, security, and societal systems at the individual and local level. Some risks within this category – particularly related to phishing, hacking, malware, or data breaches – may reach a broader scale of impact while others, such as collection of data without consent, may only impact the individual. For the most part, these risks are remediable but may include some irremediable damage.

With the rating for Inherent Probability as “Highly Likely” and Inherent Severity as “High,” the **Inherent Risk** associated with the Protection of Personal Data on the Bing service is “High.”

Risk Mitigation

Bing has implemented an expansive set of mitigations to address risks related to the Protection of Personal Data, including those described in Bing’s approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and those summarized below. The mitigations Bing has implemented relative to the Protection of Personal Data follow best practices promoting Trust and Safety in every aspect. As such, the **maturity of mitigation** efforts Bing has applied to the Protection of Personal Data risk area is assessed as “Optimized,” which brings the **Residual Risk** rating for down to “Low.”

While not a proxy for Residual Risk, Bing has measured the DDR for Protection of Personal Data on Bing Search as an average of .72% from April to June 2024. This means that Bing estimates .72% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to the Protection of Personal Data.

Nevertheless, Bing continues to invest in innovating and enhancing strategies to further mitigate and manage risks related to the Protection of Personal Data. The key mitigations currently implemented are described below.

Product Development

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative uses) to prevent Copilot from generating results containing private information hosted on low authority sites. Copilot in Bing also provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing and Image Creator from Bing and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

DPIA: For any new products or features that pose materially different personal data processing challenges, Bing is committed to completing or updating full DPIA as part of its rigorous privacy protocol. These assessments are thoroughly reviewed by Microsoft's Data Protection Officer and are subject to re-evaluation annually or sooner, should there be a significant change to processing risks. This process ensures that Bing's products and services comprehensively identify and address privacy risks at each level of development and deployment.

Privacy, Security, and Accessibility Reviews: In addition to DPIAs, Bing ensures that new features and products undergo a standard privacy and security review, which is overseen by professional compliance managers. The Privacy team evaluates products, services, and features to ensure compliance with Microsoft Privacy Standards and that personal data is protected by design and default. The Accessibility team evaluates product interfaces to ensure they are accessible to and compliant with Microsoft Accessibility Standards. The Security team evaluates products, services, and features to ensure data is secured, threats are mitigated, and compliance with Microsoft Security Standards is met. This review is critical in identifying whether any mitigations are necessary, and if so, product teams are required to complete these mitigations before any product launch. In instances where complex questions emerge regarding privacy, Bing benefits from access to cross-Microsoft privacy subject matter experts. These experts not only provide counsel on specific issues but also engage regularly with external stakeholders and regulators. Through this engagement, they address concerns and ensure that Microsoft's privacy practices align with and exceed stakeholder expectations, maintaining a steadfast commitment to user privacy and security across its operations.

Feature Evaluation: Bing takes steps to prevent the creation of, and reactively removes autosuggestions that could undermine the protection of personal data. Bing provides users with easily accessible mechanisms for reporting problematic suggestions.

User Data Controls: Users of Bing are given controls over collection and use of personal data on the service, as well as controls that allow them to exercise their data subject rights to view, access, export and delete personal data held by Microsoft. Bing maintains user data in accordance with Microsoft Privacy and Security Standards.

Microsoft Privacy Dashboard: Microsoft also provides its users with robust tools to exercise their rights over their personal data. For data that is collected by Copilot in Bing, including through user queries and prompts, the Microsoft Privacy Dashboard provides authenticated (signed-in) users with tools to exercise their data subject rights, including by providing users with the ability to view, export, and delete stored conversation history. Microsoft continues to take feedback on how they want to manage their new Bing experience, including through the use of in-context data management experiences.

- In the EU, any optional data collection or use via cookies or similar technologies (the primary mechanism for data tracking in an online service like Bing) requires opt-in consent and users can exercise their data subject rights via the Microsoft Privacy Dashboard leveraging the personalization & advertising toggle.
- In Bing's generative AI features, users can view individual initial conversation prompts through in-product features and can export and delete product and service usage data to delete stored conversation history through the Privacy Dashboard.

Chat History Feature Controls: In addition to controls available via the Microsoft Privacy Dashboard, which allow users to view, export and delete their search history, including components of their Copilot Chat history, authenticated users who have enabled the Chat history feature in the product have the ability to view, access and download chat history through in-product controls. Users may clear specific chats from Chat history or fully turn off Chat history functionality at any time by visiting the Bing Settings page. Users also may choose whether to allow personalization to access a more tailored experience with personalized answers. Users may opt-in and opt-out from personalization at any time in Chat Settings in the Bing Settings page. Clearing specific chats from Chat history prevents them from being used for personalization.

User Autonomy and Agency: Copilot in Bing will honor users' privacy choices, including those that have previously been made in Bing, such as consent for data collection and use that is requested through cookie banners and controls available in the Microsoft Privacy Dashboard. To help enable user autonomy and agency in making informed decisions, Bing has used its internal review process to carefully examine how choices are presented to users.

Product Governance

Microsoft Privacy Statement: The Microsoft Privacy Statement outlines what personal data Microsoft collects, how it is used, purposes for which it is used, and how to access and control personal data. This includes how these components apply to specific Microsoft products like search and browse, cookies, Microsoft Account information, and minor's data. Additional topics about personal data are covered such as security, storage, and retention, and how to contact Microsoft. Additionally, Microsoft offers a child-friendly version of its privacy statement to better inform younger users of data practices.

Incident Response Governance: Bing has ongoing monitoring processes to track and evaluate any deviations from privacy and security requirements or exposures of data, referred to as the incident response program. Incident response programs are governed by standard operating procedures developed with consultation from cross-Microsoft experts in global incident response and reporting requirements. Bing has ongoing monitoring processes to track and evaluate any deviations from privacy and security requirements or exposures of data, referred to as the incident response program. Incident response programs are governed by standard operating procedures developed with consultation from cross-Microsoft experts in global incident response and reporting requirements.

Safety Features: Microsoft continues to consider the needs of minors and young people as a part of the risk assessments of generative AI features in Copilot in Bing. Microsoft Accounts that identify the user as under 13 years of age or as otherwise specified under local laws cannot sign-in to access the features such as personalization that signed in users would access in Bing.

- Microsoft sets chat outputs in Copilot in Bing to Bing's SafeSearch Strict Mode, which has the highest level of safety protection in the main Bing search, hence helping to prevent users, including teen users, from being exposed to potentially harmful content. Parents have the ability to lock their children's accounts to SafeSearch Strict Mode from Windows Family Setting. In addition to information provided in this document and in FAQs regarding chat features, more information about how Copilot in Bing works to avoid responding with unexpected offensive content in search results is available here.
- Across Bing features, authenticated users under 18 are not subject to targeted advertising. This important protection for youth will be carried forward into the new Bing features as well.

Data Storage Policies: Microsoft applies industry-leading data storage policies and practices to ensure the safety and security of user data that is collected and stored. Some of these practices include, but are not limited to, storing data in Microsoft-owned and operated data centers with standard physical security access and the data is encrypted from data subject to data center in transit.

Data Retention and Deletion Policies: Copilot in Bing has data retention and deletion policies to help ensure that personal data collected through Bing's chat features is kept as long as needed.

Prohibition On Privacy Violations: Bing's terms of use governing activity on its generative AI features prohibit the use of the service to violate others' privacy. Safety systems in generative AI features are designed to prevent abuse, such as blurring faces in images before they are used as prompts in visual search and preventing the generation of images with individual faces. Bing labels images as generated by Image Creator from Bing to limit the possibility of misuse.

Copilot in Bing Data Storage Policies: Copilot in Bing was built with privacy in mind, so that personal data is collected and used as needed and is retained no longer than is necessary. As mentioned above, Visual Search in Copilot in Bing feature deploys a mechanism that blurs faces in the images at the time of the upload by users, so that facial images are not further processed or stored. More information about the personal data that Bing collects, how it is used, and how it is stored and deleted is available in the Microsoft Privacy Statement, which also provides information about Bing's new chat features.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of QC score, which is one of Bing's main ranking parameters. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

Additional Algorithmic Interventions: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or "defensive search interventions," to counteract these threats, in accordance with its search principles.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing's content removal practices for the search index are largely reactive to reports and involve human review. Bing's support teams are provided with training and oversight in order to ensure removal practices align with Bing's principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing. Users may report content that violates Bing policies prohibiting the indexing of private content published without consent such as credit cards, or other private information, or images of minor users. The content review teams are well-trained to efficiently handle such requests, and Bing conducts regular system tests and audits to maintain its responsiveness.

Honoring the Right to be Forgotten: Bing's established "Right to Be Forgotten" policies and procedures help reduce risks of negative effects to protection of personal data by providing reporting tools for users to submit requests that certain online content associated with them be removed from Bing. Consistent with the EU Right to be Forgotten, EU users may submit requests to remove search results for queries that include the person's name where they can demonstrate the results are inadequate, inaccurate, no longer relevant, or excessive using Bing's Request Form to Block Search Results in Europe. Bing has dedicated teams, training materials, and escalation paths for "Right to Be Forgotten" request and removes indexed websites (and, as relevant, search suggestions) where the requestor or data subject's privacy interest is not outweighed by the public interest. Bing also engages with local courts and Data Protection Authorities on data subject escalations of Right to Be Forgotten decisions and responds to feedback as appropriate.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of harm and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

Data Retention: Data retention is managed within a set schedule. Bing limits the retention of data to that necessary to provide the service and ensure the safety of the service.

Online Data Store: Data-related to Bing's search service is stored in Microsoft-owned data centers, maintaining standard physical security and encryption during transit. The **Online Data Store**, which improves service interaction through stored preferences, is geographically distributed to enhance performance. Access to most Bing search history data is limited to Bing employees.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing's generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing's product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing's operations.

Suspension of Microsoft Advertising Services: Similarly, Bing reserves the right to issue immediate suspensions of advertiser's accounts should they egregiously violate Microsoft's terms or policies, if those violations intend to use Microsoft services in a fraudulent, harmful, or illegal manner. Microsoft uses both manual and systematic methods to review ads for compliance with their policies and reserves the right to remove those that are non-compliant.

Product Improvement

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of negative impacts to the Protection of Personal Data on Bing's services.

External Collaboration: Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to help ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform the processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or other special information for risks such as malicious websites that could seek user data.

Content Provenance and Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Transparency and Controls: To unlock the transformative potential of generative AI, Microsoft believes it must build trust in the technology through empowering individuals to understand how their data is used and providing them with meaningful choices and controls over their data. Copilot in Bing is designed to prioritize human agency, through providing information on how the product works as well as its limitations, and through extending robust consumer choices and controls to Copilot in Bing features.

Privacy Statement: The Microsoft Privacy Statement provides information about transparent privacy practices for protecting customers, and it sets out information on the controls that give users the ability to view and manage their personal data. To help ensure that users have the information they need when they are interacting with Bing's new conversational features, in-product disclosures inform users that they are engaging with an AI product, and Microsoft provides links to further FAQs and explanations about how these features work. Microsoft will continue to listen to user feedback and will add further detail on Bing's conversational features as appropriate to support understanding of the way the product works.

Content Removal: Bing is transparent about its privacy practices and content removal policies, as well as reports on content removed due to privacy concerns such as Right to be Forgotten.

Illegal Content and Activities

Risks related to Illegal Content and Activities include potential intellectual property infringement and defamation, and the promotion or sale of illegal, dangerous, or regulated goods or services. Bing's content moderation systems are designed to process requests regarding illegal content takedown to be reviewed by appropriate teams. Illegal content removal actions are tracked and disclosed in Bing's transparency reports. Bing takes proactive steps to prevent the dissemination of CSEAI using PhotoDNA scanning. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Illegal Content and Activities, the Bing Risk Assessment team has assessed the Residual Risk related to Illegal Content an Activities on the service to be Low.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Illegal Content and Activities	Likely	Critical	High	Managed	Low

Risk Definition

Risk Area	**Illegal Content**
Risk Definition	Risk related to the conduct of illegal activity or dissemination of illegal content through the service.
Theoretical Risk Manifestations absent Mitigations	• Risk that Search results return results containing illegal content • Risk that Bing Shopping or search results include websites related to the sale of illegal, dangerous, infringing, or regulated goods (including illegal drugs, weapons, psychoactive substances, pharmaceuticals, etc.) • Risk that Copilot in Bing or Image Creator from Bing are used to generate third party IP without authorization in generated responses • Risk that Advertisements promote illegal content • Risk that generative AI features generate content that violates local laws • Risk that Search suggestions suggest queries that might lead to illegal content sources, such as third-party websites offering illegal content streaming or music downloads

Risk Analysis

The **inherent probability** of systemic risks related to Illegal Content appearing on deriving from use of Bing's products and services absent sufficient mitigations is assessed as "Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a “Likely” level of probability relative to other potential systemic risks on Bing.
- Due to the nature of Bing as a search engine, Bing’s goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing may provide relevant results even if they are less authoritative, while working to ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2%

use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** of impact for content and activities that negatively impact "Illegal" Content and Activities on Bing is rated as "Critical" primarily due to the gravity of risks within this category considering the potential for harm to security, environment, and wellbeing systems up to the regional and even the global level for environment systems. Some risks within this category have potentially irremediable consequences.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with Illegal Content and Activities on the Bing service is "High."

Risk Mitigation

Bing has implemented a robust set of mitigations to address risks related to Illegal Content and Activities, including those described in Bing's approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and those summarized below. The mitigations Bing has implemented relative to Illegal Content and Activities follow best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts Bing has applied to the Illegal Content and Activities risk area is assessed as "Managed," which brings the **Residual Risk** rating for down to "Low."

While not a proxy for Residual Risk, Bing has measured the DDR for Illegal Content and Activities on Bing Search as an average of .83% from April to June 2024. This means that Bing estimates .83% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Illegal Content and Activities.

Nevertheless, Bing continues to invest in developing and enhancing strategies to further mitigate and manage risks related to Illegal Content and Activities. The key mitigations currently implemented are described below.

Product Development

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing and Image Creator from Bing and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

New Feature Launch: Bing relies on Microsoft's extensive cross-company compliance infrastructure to proactively review new products and features to ensure they meet Microsoft's policy commitments in key areas such as privacy, accessibility, digital safety, and RAI.

Product Governance

Algorithmic Mitigations: Bing has invested in robust principles and algorithmic mitigation capabilities to effectively mitigate the risk to dissemination of "illegal" content on the service, which includes IP infringement; defamation; and promotion or sale of illegal, dangerous, or counterfeit goods, services. Bing deploys proactive automated detection for CSEAI and removes confirmed CSEAI based on Microsoft policy. Bing employs additional safeguards to ensure any actions taken are narrow, specific, submitted in writing, and based on valid legal orders for any orders received from the government.

Terms & Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft Services Agreement Code of Conduct broadly prohibits using Microsoft services to generate or share content that is illegal. In addition to the Microsoft Services Agreement, Users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with law-violating content, including unauthorized IP generation, illegal streaming or downloading, and sale of illegal items).

Abuse/Spam Policies: Bing's general abuse/spam policies, detailed in Bing's Webmaster Guidelines, include details regarding prohibiting certain practices intended to manipulate or deceive the Bing search algorithms.

Partnership with Internal Subject Matter Experts: The Bing team also relies on Microsoft's extensive team of subject matter experts on global regulatory issues to ensure awareness of new obligations that may arise in markets where it operates and as well as specialized legal subject matter expert teams focused on intellectual property, privacy, accessibility, security, digital safety, and responsible AI.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of QC score, which is one of Bing's main ranking parameters. Generally, websites that appear dedicated primarily toward offering illegal content (such as pirated music files) are considered as low authority and ranked accordingly. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

- Bing users generally cannot share their own content with the broader user base on the service, so users cannot disseminate illegal materials through the service. However, third-party website content linked in search results may include illegal materials. Bing’s ranking algorithms are designed to help users find the most relevant, highest authority content available in response to their search queries (thus limiting the appearance of illegal materials).

Additional Algorithmic Interventions: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or “defensive search interventions,” to counteract these threats, in accordance with its search principles.

- Bing works to prevent the display of autosuggestions that could lead users to illegal materials and demotes websites in search results that receive a large volume of copyright removal notices. Bing has a dedicated team that is accountable for implementing algorithmic interventions and metrics monitoring and dedicated to identifying high-risk areas where Bing's ranking and relevance algorithms deviate from its principles and goals, through methods such as red team testing, external threat intelligence, and social listening systems. Bing also adds additional information to avoid misleading users such as PSAs and warnings on illegal materials and help ensure these interventions are available across the markets and languages in which Bing is offered.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing’s content removal practices for the search index are largely reactive to reports and involve human review. Bing’s support teams are provided with training and oversight to ensure removal practices align with Bing’s principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

- Given that third-party website content linked in search results may include illegal materials, Bing has robust content moderation principles to guide the moderation of web results to limit the amount of illegal content appearing on the service.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing’s generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing’s product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing’s operations.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing. The Report a Concern forms include functionality to report unlawful content.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of harm and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

Content Removal and Suspension: For content removal, Bing will remove materials identified as illegal on notice, and in addition scans for CSEAI using **PhotoDNA**. Bing also takes steps to globally remove (based on Bing policies) certain categories of content that may be illegal materials in some jurisdictions, such as extremely sensitive personal information that could be used for identity theft or other fraud. Similarly, Bing also states in their policies that advertisers who are misusing Microsoft services in a way that could cause harm to users or the service, such as advertising in fraudulent, harmful, or illegal manner are subject to immediate suspension.

Suspension of Microsoft Advertising Services: Microsoft uses both manual and systematic methods to review ads for compliance with their policies and reserves the right to remove those that are non-compliant.

Product Improvement

External Collaboration: Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to help ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform the processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Representative examples of external engagements undertaken during the Reporting Period relevant to Illegal Content include: Global Internet Forum to Counter Terrorism (GIFCT), the WeProtect Global Alliance, the Christchurch Call to Action, the Global Project Against Hate and Extremism, the United Nations' International Narcotics Control Board, the Anti-Defamation League, and the EU Council's "Horizontal Working Party on Enhancing Resilience and Countering Hybrid Threats" ("ERCHT"), among others.

Bing's commitment to digital safety is also evident in its proactive approach to countering the exploitation of online services, particularly in the fight against the trafficking of dangerous synthetic opioids. In March 2024, Bing highlighted its practical approaches to mitigating risks associated with cross-platform exploitation during a UN Internation Narcotics Control Board (INCB) event. Bing's search principles are designed to deliver credible and authoritative results, ensuring that users have access to information that is both free and lawful, while also respecting local laws and fundamental rights. Bing takes significant steps to shield users from harmful and offensive content, maintaining transparency about its principles, practices, and actions.

- To further bolster the fight against illegal content, Bing leverages both internal and authoritative external sources, such as the IWF and the International Social Service (ISS), to enhance its ranking and relevance capabilities. This approach is particularly crucial in identifying and demoting potentially illegal materials, where, for instance, a high volume of valid copyright infringement notices can signal a site's low quality, influencing its ranking negatively.
- Bing has recently taken significant steps to bolster its digital safety measures, engaging in discussions with two major Very Large Online Platforms (VLOP) to explore a collaborative approach to combat the proliferation of synthetic drug trafficking online. This initiative is part of a broader effort that includes (possible) government-to-industry information sharing.

Algorithmic Prioritization of High Authority Content: To provide users with the most relevant and highest authority content in search queries, Bing invests significant time and resources into ranking and relevance systems. Bing regularly reviews product quality issues, recurring trends, and emerging risks to ensure its algorithms are sufficient without subjecting the service to unnecessary censorship. In most cases, the user will be able to find the content in standard search results regardless of content being removed from generative features. By detailing the core parameters for content ranking in its Webmaster Guidelines and conducting rigorous evaluations of its systems through objective metrics, Bing not only strives to limit the visibility of illegal or harmful materials but also maintains a transparent and reliable search environment.

- Bing's ongoing efforts include continuous improvements to its enhanced search features and additional safety guardrails are implemented to further secure the service. Content moderation is a key focus, with Bing reviewing user reports and legal takedown requests diligently. Digital literacy initiatives are in place to educate and empower users, and strict advertising policies are enforced to prevent illegal activities and the promotion of dangerous substances.
- Additionally, Bing has proactively expanded its defensive algorithms by incorporating over 2,000 new terms related to non-scheduled opioids, along with their chemical numbers and coded terms.
- This dedicated effort is supported by a significant team focused on ongoing improvements in ranking and relevance, demonstrating Bing's comprehensive approach to mitigating the risk of illegal content and activities on the service.

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of Illegal Content and Activities occurring on Bing's services.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or other special information "hubs" related to public health issues.

Content Provenance & Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Notice to Users: Bing takes a proactive stance towards enhancing product transparency and mitigating the risks associated with illegal content and activities on its service. Through the implementation of several measures, Bing aims to inform and protect its users from potential harm. This includes the integration of PSAs and warnings in search results that are likely to lead to illicit or harmful materials, such as illegal pharmaceuticals or websites hosting malware.

Microsoft Adverting Complaint Intakes: Users are encouraged to report advertisements that promote illegal content or violate Microsoft Advertising terms through a structured feedback form available on Bing's pages. Specific forms for reporting low quality ads or IP concerns are also available.

Private and Family Life

Risks related to Private and Family Life include exposure to content pertaining to the malicious sharing or exploitation of private data, doxing, disclosure of private images, privacy intrusions. Bing's principles allow for removal of third-party search results that have an undue impact on privacy, such as content that can be used for ID theft or fraud, sensitive content that was inadvertently disclosed such as emails or health records, and nonconsensual pornography. In the EU, Bing respects the Right to be Forgotten. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Private and Family Life, the Bing Risk Assessment team has assessed the Residual Risk related to Private and Family Life on the service to be Low.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Private and Family Life	Likely	High	High	Managed	Low

Risk Definition

Risk Area	Private and Family Life
Risk Definition	Risk that content or activities with actual or foreseeable negative effect on respect for private and family life occur on the service.
Theoretical Risk Manifestations absent Mitigations	• Risk that Search results include NCII • Risk that bad actors misuse Copilot in Bing to accelerate the creation of content that could facilitate doxing, exposure of private or intimate information, or identity theft • Risk that Image Creator from Bing could be misused to generate synthetic, yet realistic private or intimate images of individuals • Risk that street-level imaging features in Maps may inadvertently compromise individuals' privacy • Risk that practices of Advertisements on Bing, particularly those involving targeted or personalized ads, intrude upon individuals' private lives by making inferences or collecting data without explicit consent, potentially leading to unwelcome privacy intrusions and undermining individuals' autonomy over their personal information

Risk Analysis

The **Inherent Probability** of systemic risks related to Private and Family Life stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a “Likely” level of probability relative to other potential systemic risks on Bing.
- User intent is generally required for users to access potentially harmful content on Bing. Due to the nature of Bing as a search engine, Bing’s goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less credible, while working to ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** of impact for content and activities that negatively impact Private and Family Life on Bing is rated as “High” primarily due to the gravity of risks within this category, considering the potential for harm to economic, security, societal, and wellbeing systems at the individual-level with some broader reverberations. Some risks in this category, such as posts that contain sensitive data or disclosure of nonconsensual private or intimate images, have potentially irremediable impacts. Some risks in this category, such as users' inadvertent sharing of sensitive personal information (e.g., GDPR Article 9 information that includes racial or ethnic origin, political opinions, religious or philosophical beliefs etc.) may be less severe.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "High," the **Inherent Risk** associated with Private and Family Life on the Bing service is "High."

Risk Mitigation

Bing has implemented a robust set of mitigations to address risks related to Private and Family Life, including those described in Bing's approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and those summarized below. The mitigations Bing has implemented relative to Private and Family Life follow best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts Bing has applied to the Private and Family Life risk area is assessed as "Managed," which brings the **Residual Risk** rating for down to "Low."

While not a proxy for Residual Risk, Bing has measured the DDR for Private and Family Life on Bing Search as an average of .74% from April to June 2024. This means that Bing estimates .74% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Private and Family Life.

Nevertheless, Bing continues to invest in developing and enhancing strategies to further mitigate and manage risks related to Private and Family Life. The key mitigations currently implemented are described below.

Product Development

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

- Bing is committed to conducting thorough RAI reviews, especially concerning the implementation of generative AI features. This initiative is geared towards minimizing any possible harm that could arise from the use of advanced AI technologies. By rigorously evaluating these technologies before they are introduced to the platform, Bing aims to preemptively address any negative impact of artificial intelligence on the privacy and family life of its users. This dual strategy underscores Bing's dedication to maintaining a safe and respectful online environment, where the rights to privacy and family life are protected against the evolving landscape of digital threats.

New Feature Launch: Bing works to ensure that any new product or feature does not unduly impact privacy and conducts privacy, security, and accessibility reviews in its product design process to minimize this risk. Privacy, security, and accessibility reviews are an essential part of any new feature review process; implementing recommended privacy mitigations following such reviews is a requirement for launch. Microsoft has a robust privacy and security infrastructure, consisting of privacy managers who are trained in Microsoft privacy standards and relevant laws, and have access to centralized privacy specialist teams and legal support for complex or novel issues.

Privacy, Security, and Accessibility Reviews: The Privacy team evaluates products, services, and features to ensure compliance with Microsoft Privacy Standards and that personal data is protected by design and default. The Accessibility team evaluates product interfaces to ensure they are accessible to and compliant with Microsoft Accessibility Standards. The Security team evaluates products, services, and features to ensure data is secured, threats are mitigated, and compliance with Microsoft Security Standards is met.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing and Image Creator from Bing and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

Privacy and Security Infrastructure: Bing has a robust privacy and security infrastructure that includes full-time professional compliance managers, requires pre-launch feature reviews and mitigations, adherence to strict standards for data handling and security by internal employees and vendors, training for employees, and completion of DPIA ensure appropriate risk mitigations.

User Data Controls: Users of Bing are given controls over collection and use of personal data on the service, as well as controls that allow them to exercise their data subject rights to view, access, export and delete personal data held by Microsoft. Bing maintains user data in accordance with Microsoft security standards. Bing has instituted a series of measures aimed at mitigating the risks associated with third-party content that could potentially negatively impact EU users' rights to private and family life.

Product Governance

Terms & Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft Services Agreement Code of Conduct broadly prohibits using Microsoft services to violate or infringe upon the privacy and rights of others. In addition to the Microsoft Services Agreement, Users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with intrusive activities, including unauthorized sharing of copyrighted material, or sharing content of others without their consent).

User Input Reviews: Bing regularly reviews user complaints via formal reporting channels as well as the user Feedback portal, which informs updates to principles and procedures. Bing does not allow users to post or share content on the service but does use this feedback on content appearing in search results and generative AI features to inform its principles and practices. See User Feedback and Reporting section of the Appendix II.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of QC score, which is one of Bing's main ranking parameters. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

Additional Algorithmic Interventions: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or "defensive search interventions," to counteract these threats, in accordance with its search principles.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of harm and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing. Bing's Report a Concern tool includes options for users to report exposed personal or private information as well as non-consensual intimate imagery.

Honoring the Right to be Forgotten: Bing’s established “Right to Be Forgotten” policies and procedures help reduce risks of negative effects to private and family life by providing reporting tools for users to submit requests that certain online content associated with them be removed from Bing. Consistent with the EU Right to be Forgotten, EU users may submit requests to remove search results for queries that include the person's name where they can demonstrate the results are inadequate, inaccurate, no longer relevant, or excessive using Bing’s Request Form to Block Search Results in Europe. Bing has dedicated teams, training materials, and escalation paths for “Right to Be Forgotten” request and removes indexed websites (and, as relevant, search suggestions) where the requestor or data subject’s privacy interest is not outweighed by the public interest. Bing also engages with local courts and Data Protection Authorities on data subject escalations of Right to Be Forgotten decisions and responds to feedback as appropriate.

Content Removal: Moreover, Bing prioritizes the balance between information access and privacy. Content removal is carefully considered by Bing, especially in scenarios where privacy concerns outweigh the public interest in access to information. Bing's content policies cater to the removal of third-party search results that pose undue privacy risks, including instances of identity theft, fraud, and nonconsensual pornography reported to Bing. The service also upholds the Right to be Forgotten for EU users, removing private content where appropriate and demonstrating a respect for individual privacy.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing’s generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing’s product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing’s operations.

Facial Blurring: Bing visual search features (which allow use of images as a search prompt) do not allow for matching of images of private individuals’ faces with content in the search index. It also does not allow for searching of adult images. In the visual search feature in Copilot in Bing, Bing blurs faces before processing data contained in files provided by the user.

Privacy Training: Bing has implemented a multifaceted approach to product enforcement to protect the right to private and family life across its service. Central to this strategy is the requirement for Microsoft personnel to undergo annual privacy training. This ensures that each team member is well-versed in appropriate data use practices, reinforcing the company's commitment to safeguarding user privacy.

Third-Party Engagements: In its external relations, Microsoft, including Bing, engages in third-party evaluations by the independent organization Ranking Digital Rights, which assesses Bing’s practices, governance, and leadership in protecting privacy.

Problematic Suggestions: Bing takes steps to prevent the creation of suggestions that could undermine the right to private and family life and provides users with easy mechanisms to report problematic suggestions for removal.

Product Improvement

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of negative impacts to Private and Family Life on Bing's services.

External Collaboration: Both Bing and specialized cross-Microsoft teams that support broadly across the company regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or other special information "hubs" related to public health issues.

Content Provenance & Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Audit Reports: Bing engages in regular self-directed tests and audits of its system to ensure its ability to respond and Microsoft, including Bing, also responds to periodic evaluations by the third-party independent organization **Ranking Digital Rights** which publishes annual ratings on Bing's practices, governance, and leadership on the protection of freedom of expression and privacy.

Discrimination and Hate

Risks related to Discrimination and Hate include exposure to content containing bias, discriminatory content, and hate speech. Bing applies its comprehensive safety mitigations to the wide range of web content that poses risks to discrimination and hate. By prioritizing high authority content, Bing's ranking algorithms are designed to protect users from being unexpectedly harmed by biases, hate speech, and other discriminatory content. Generative AI features deploy additional mitigations to prevent harmful AI-generated outputs. Additionally, generative AI features undergo review prior to launch to identify the potential for harm, including harm related to Discrimination and Hate, under Microsoft's RAI program. Bing also provides users with SafeSearch option to control their search experience. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Discrimination and Hate, the Bing Risk Assessment team has assessed the Residual Risk related to Discrimination and Hate on the service to be Low.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Discrimination and Hate	Likely	High	High	Managed	Low

Risk Definition

Risk Area	Discrimination and Hate
Risk Definition	Risk that content or activities with actual or foreseeable negative effect on the fundamental right to non-discrimination occur on the service.
Theoretical Risk Manifestations absent Mitigations	• Risk that Copilot in Bing results or suggested prompts reflect, reinforce, or perpetuate stereotypes, biases, or inequalities in the content generated • Risk that Bing algorithmic systems contain bias that negatively impacts protected user groups on Bing • Risk that users leverage Image Creator from Bing or Copilot in Bing to create discriminatory or hateful content • Risk that third-party websites contained in Search results contain hate speech or discriminatory content • Risk that safety features and interventions are designed or implemented without adequate linguistic or cultural understanding, resulting in impacts to some users because of protected traits • Risk that advertisements on Bing are biased in their targeting, affecting protected groups' ability to access critical services

Risk Analysis

The **Inherent Probability** of systemic risks related to Discrimination and Hate stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a “Likely” level of probability relative to other potential systemic risks on Bing.
- User intent is generally required for users to access potentially harmful content on Bing. Due to the nature of Bing as a search engine, Bing’s goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less credible, while working to ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2%

use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** for content and activities that negatively impact Discrimination and Hate on Bing is rated as "High" primarily due to the gravity of risks within this category, considering the impact to wellbeing, societal, and economic impacts at the individual-level and societal impacts up to a regional scale. Risks within this category, such as hate speech, and discrimination may be remediable but can cause serious damage including loss of career advancement opportunities.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "High," the **Inherent Risk** associated with Discrimination and Hate on the Bing service is "High."

Risk Mitigation

Bing has implemented a robust set of mitigations to address risks related to Discrimination and Hate, including those described in Bing's approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and those summarized below. The mitigations Bing has implemented relative to Discrimination and Hate follow best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts Bing has applied to the Discrimination and Hate risk area is assessed as "Managed," which brings the **Residual Risk** rating for down to "Low."

While not a proxy for Residual Risk, Bing has measured the DDR for Discrimination and Hate on Bing Search as an average of .76% from April to June 2024. This means that Bing estimates .76% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Discrimination and Hate.

Nevertheless, Bing continues to invest in developing and enhancing strategies to further mitigate and manage risks related to Discrimination and Hate. The key mitigations currently implemented are described below.

Product Development

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing and Image Creator from Bing and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

New Feature Launch: Bing relies on Microsoft's extensive cross-company compliance infrastructure to proactively review new products and features to ensure they meet Microsoft's policy commitments in key areas such as privacy, accessibility, digital safety, and RAI.

Product Governance

Terms & Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft Services Agreement Code of Conduct broadly prohibits using Microsoft services to engage in activity that is harmful to others. In addition to the Microsoft Services Agreement, Users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with harmful content, including hate speech).

Image Creator from Bing Policies: Bing enforces a set of policies specifically for Image Creator from Bing, including the RAI Harm Guidelines, which provides a harms taxonomy for identifying harmful prompts or prompts that may generate harmful content. The Guidelines also provide RAI metrics for measuring Image Creator from Bing's effectiveness in preventing the generation of discriminatory or harmful content.

Microsoft Advertising Policies: Microsoft's Advertising policies apply to ads on Bing. Microsoft removes Ads that violate prohibitions on biased and discriminatory content.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of QC score, which is one of Bing's main ranking parameters. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

Additional Algorithmic Interventions: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or "defensive search interventions," to counteract these threats, in accordance with its search principles.

- Bing's algorithms, which focus on promoting high authority sources, are designed to reduce the likelihood of discriminatory or hateful content in higher ranked search results (in cases where this type of content is not the intended result of the user). Bing deploys targeted algorithmic interventions to reduce the prevalence of this risk, while enforcement teams take action on users creating harmful content pertaining to Discrimination and Hate using Copilot in Bing or Image Creator from Bing.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing's content removal practices for the search index are largely reactive to reports and involve human review. Bing's support teams are provided with training and oversight to ensure removal practices align with Bing's principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing's generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing's product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing's operations.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of harm and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's

AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

- Regarding Bing's generative AI features, the same mitigations are employed with additional enhanced safety features such as classifiers, filters, and a bespoke metaprompt that further limit the likelihood of harmful content appearing in Copilot in Bing/Image Creator from Bing features. Bing has engaged in extensive RAI reviews regarding generative AI features to ensure outputs are not biased or discriminatory.

Search Results Ranking: Bing's primary approach to reducing the visibility of Discrimination and Hate in search results focuses on promoting high authority sources in search results. As a part of Bing's commitment to a safe online environment and goal to ensure that users can access high quality and authoritative content online, Bing's ranking principles consider low quality content to include sources that use offensive statements, derogatory language, or name-calling to be low quality.

Dedicated Expert Teams: Bing employs a specific team that is dedicated to deploying targeted algorithmic interventions to reduce the prevalence of hateful speech, harassment against protected groups and other problematic content in search. Bing maintains a set of metrics to track, monitor and review the efficacy of its interventions on an ongoing basis to ensure that they are performing as expected and not inadvertently introducing additional bias or other harms. Please see the Appendix section on Abuse Pattern Analysis for more information on.

Product Improvement

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of Discrimination and Hate on Bing's services.

External Engagement and Collaboration: Both Bing and specialized cross-Microsoft teams that support broadly across the company regularly engage with external stakeholders in areas of key policy priorities, including discrimination and hate. For example, Bing has engaged Global Project Against Hate and Extremism Project on issues related to hate speech. Bing (as part of Microsoft's broader engagement) collaborates with the Anti-Defamation League (ADL) to participate in roundtable discussions on topics relating to Discrimination and Hate, with a focus on how content policy and enforcement mitigate the risk of harm in content and generative AI, and a more recent, specific focus on reducing antisemitism.

- As part of this engagement, Bing has also evaluated Copilot in Bing's factual accuracy and sensitivity relating to political debate and real-world events.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or other special information "hubs" related to public health issues.

Content Provenance & Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Notice To Users: Bing implements PSAs or warnings to inform users of harmful content relating to Discrimination and Hate. Bing's approach emphasizes the importance of making these mitigations accessible across the diverse markets and languages it serves, demonstrating its commitment to fostering an inclusive and respectful online environment.

Civic Discourse and Electoral Processes

Risks related to Civic Discourse and Electoral Process include exposure to content containing election misinformation, risks third party content indexed in search or generated in AI features could lead to political polarization, and targeted efforts by bad actors to leverage technology to undermine civic institutions and democratic elections.

Search engines like Bing play a key role in ensuring users have access to information related to civic discourse and electoral processes. Balancing user safety with freedom of information and expression and avoiding the introduction of bias or influence is critical to this risk area. Bing takes this responsibility seriously and has implemented robust measures to promote election integrity and reduce the risk of systemic harm related to civic discourse and electoral processes on Bing.

Global elections are highly complex, as each country has different election structures, voting processes, timing, political systems, and party structures. During the Reporting Period, there were an unprecedented number of key elections globally, including the EU Parliament Elections and French legislative elections. Mitigating risks related to Civic Discourse and Electoral Processes on the Bing service presents challenges in the scale and speed of electoral process, developments surrounding elections, parties, and candidates, and threats posed by malicious actors. The scale of the 2024 global elections requires simultaneous execution of election protections across multiple markets and languages, drawing up local, regional, and national context, and robust incident response systems.

This complexity is furthered by the rapid rise in generative AI accessible online. In recognition of these novel risks and harms, in November 2023 Microsoft announced a set of Election Protection Commitments grounded in four principles to help safeguard voters, candidates and campaigns, and election authorities worldwide. These principles, which help inform Bing's response and safety program, are:

- Voters have a right to transparent and authoritative information regarding elections.
- Candidates should be able to assert when content originates from their campaign and have recourse when their likeness or content is distorted by AI for the purpose of deceiving the public during the course of an election.
- Political campaigns should protect themselves from cyber threats and be able to navigate AI with access to affordable and easily deployed tools, trainings, and support.
- Election authorities should be able to ensure a secure and resilient election process and have access to tools and services that enable this process.

Bing takes a multifaceted approach to protecting civic discourse and electoral processes and regularly updates its policies, and practices to adapt to evolving risks, trends, and technological innovations, as well as regulatory expectations. Bing partners closely with dedicated Microsoft internal teams focused on elections, including Democracy Forward and specialized cross-functional teams dedicated to elections and specific risks arising from AI & Elections. In addition to Bing's core ranking algorithms and systems designed to promote high authority content, Bing deploys numerous product and process improvements consistent with DSA Election Guidelines, including ingestion of election and misinformation-related threat intelligence narratives and localized election data to inform product interventions, direction of users to official sources and high authority content, partnerships with internal and external experts to collect insights on emerging threats and trends to inform safety efforts, display of special "How to Vote" answers for EU users, rapid incident response processes, coordination (through Microsoft) with election authorities, and investment in generative AI mitigations and improvements intended to address risks of hallucination or inaccurate information related to elections and political deepfakes. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Civic Discourse and Electoral Processes, the Bing Risk Assessment team has assessed the Residual Risk related to Civic Discourse and Electoral Processes on the service to be Low.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Civic Discourse and Electoral Processes	Likely	Critical	High	Optimized	Low

Risk Definition

Risk Area	Civic Discourse and Electoral Processes
Risk Definition	Risk that content or activities with actual or foreseeable negative effects on civic discourse and electoral processes occur on the service.
Theoretical Risk Manifestations absent Mitigations	• Risk that bad actors misuse Image Creator from Bing to generate deceptive election content, such as deceptive images of political candidates, elections, and voting processes • Risk that Search results or Search suggestions promote low authority, low quality, outdated, or other content that includes inaccurate, misleading information about electoral processes, voting procedures, or election outcomes in EU member states and around the world • Risk that lack of authoritative information or data voids in less trafficked languages or geographies leads to lower quality search results in certain markets • Risk that political advertisements target vulnerable Bing users with election-related misinformation • Risk that Copilot in Bing responses link to low authority, low quality, or outdated content with inaccurate or misleading information about elections or electoral processes • Risk that Copilot in Bing mischaracterizes cited web links or otherwise provides inaccurate information about elections or electoral processes (e.g., hallucinations, inconsistency, or grounding issues) • Risk that News promotes low authority or low quality news sources that increase political polarization and filter bubbles

	• Risk that Bing’s content moderation policies or processes and safety systems, introduce bias or limit access to important election information if applied inconsistently • Risk that Bing’s partnerships with external experts or fact checkers introduce bias into the Bing service • Risk that content promoting efforts to undermine election integrity, or mass activities of groups seeking to destabilize civic institutions (e.g., anti-democracy organizations) appears in Search results or Copilot in Bing outputs

Risk Analysis

The **Inherent Probability** of systemic risks related to Civic Discourse and Electoral Processes stemming from the use, misuse, or functioning of the Bing service absent sufficient mitigations is assessed "Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a “Likely” level of probability relative to other potential systemic risks on Bing.
- Due to the nature of Bing as a search engine, Bing’s goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less credible, while working to ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.
- The Bing Risk Assessment team also considered a high number of elections in Europe over the risk assessment period as a potential impact to the probability level.
- The risk of probability is also informed by actual usage. During EU Parliament Elections, for example, Bing estimated that only approximately 0.2-0.5% of prompts in Copilot in Bing seek election-related information. While these estimates cannot necessarily capture all user intent, they do provide insight into how users use generative AI in the context of finding election-related information, such as how or where to vote.

The **Inherent Severity** for content and activities that negatively impact Civic Discourse and Electoral Processes on Bing is rated as “Critical” primarily due to the gravity of risks within this category, considering the potential for significant harm to political, societal, economic, and security systems at the local, country, and even regional levels. Some risks within this category have potentially irremediable consequences on political or security systems. The impact of other risks may be considered remediable but still has the potential to cause significant damage.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with Civic Discourse and Electoral Processes on the Bing service is "High."

Risk Mitigation

Bing has implemented an expansive set of mitigations to address risks related to Civic Discourse and Electoral Processes, including those described in Bing's approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and those summarized below. The mitigations Bing has implemented relative to Civic Discourse and Electoral Processes follow best practices promoting Trust and Safety in every aspect. As such, the **maturity of mitigation** efforts Bing has applied to the Civic Discourse and Electoral Processes risk area is assessed as "Optimized," which brings the **Residual Risk** rating for down to "Low."

In addition, Bing has reviewed the guidance provided by the European Commission (Commission Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes pursuant to Article 35(3) of Regulation (EU) 2022/2065). While many of Bing's existing practices are reflected in the guidelines, Bing made additional improvements to its risk mitigations for elections consistent with guidelines suggested by the Commission and Bing and Microsoft have undertaken multiple engagements with the Commission, Digital Service Coordinators, and election authorities in the Union to understand concerns and share information about Bing's risk mitigation strategies in the lead up to critical EU elections.

While not a proxy for Residual Risk, Bing has measured the DDR for Civic Discourse and Electoral Processes on Bing Search as an average of .71% from April to June 2024. This means that Bing estimates .71% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Civic Discourse and Electoral Processes.

Nevertheless, Bing continues to invest in innovating and enhancing strategies to further mitigate and manage risks related to Civic Discourse and Electoral Processes. The key mitigations currently implemented are described below.

Product Development

Algorithmic Prioritization of High Authority Content: Bing's primary mechanism for combatting behaviors in search results that could negatively impact electoral processes is via its ranking algorithms and systems designed to identify and combat attempts to abuse search engine optimization techniques (i.e., spam). How Bing Delivers Search Results and the Bing Webmaster Guidelines contain Bing's principles for ranking and moderation of third-party content in web results and provide detailed information on the removal of content that violates laws or Bing principles.

Regular Review of Election Integrity Features and Strategy: Bing takes a multi-layered approach to protecting election integrity and regularly updates its strategies and practices to adapt to evolving risks, trends, and technological innovations, as well as new election announcements (many of which occur on short notice) and regulatory guidance.

Democracy Forward Program: Microsoft established its Democracy Forward team in 2018 to lead the company's collective efforts to help safeguard elections and democratic institutions around the world, including the EU. The initiative consists of numerous programs designed to help protect the integrity of electoral processes and promote the security of elections. Bing works closely with Democracy Forward on election-related features, policies, and mitigations as well as to collect regional and local context on elections. Through Democracy Forward, Microsoft engages in partnerships across industry and with civil society, including the International Foundation for Electoral Systems (IFES), International IDEA, NATO Hybrid Center of Excellence among others, on a consistent basis to identify and promote best practices related to election integrity.

Dedicated Internal Teams: In addition to Democracy Forward, there are several cross-functional teams at Microsoft dedicated expressly to addressing election related issues. Microsoft has undertaken efforts to improve its processes to support election integrity using a whole-of-company approach. The company has dedicated internal teams focused on elections including Democracy Forward and specialized cross-functional teams established specifically to work on issues pertaining to the 2024 election year. This includes review of election-related product features and a specialized team dedicated to addressing risks, policies, and issues surrounding AI and Elections, all of which feed into product decisions and design for Bing search, generative AI features, and other applicable products. For example, Microsoft established a 2024 election year working group that brings together key stakeholders from across the Bing and Copilot teams, local support, and broader Microsoft policy professionals to discuss a broad array of elections-related issues ranging from responsible AI and election security to planned product improvements. This group meets regularly to discuss existing, planned, and proposed features and mitigations, to share learnings and new information, and to highlight existing and emerging areas of risk with respect to elections and civic integrity around the world. In addition, there are product-specific election teams tasked with implementing election features appearing in Bing, global elections working groups, and a specialized EU elections working group at Microsoft.

EU Code of Practice on Disinformation (“COPD”): As a signatory to the COPD and active member of COPD working groups, Bing complies with commitments in the code dedicated to reducing misinformation risks, including participation in the Elections Working Group rapid response program, providing detailed transparency reporting, providing users with tools and functionality to help understand the trustworthiness of the sites or domains they are visiting and Bing’s empowering them to make informed decisions about those sources, and supporting good faith research into disinformation, among other commitments. Bing also provide enhanced reporting on Elections as a designated “crisis” under COPD.

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch. Microsoft’s Office of Responsible AI has established policies concerning election-related content and works with teams, including Bing Search, Copilot in Bing, and Image Creator from Bing on implementation.

Grounding GenAI in High-Ranking Authority Search Results: Copilot in Bing textual responses are generally grounded in web search results. “Grounding” refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

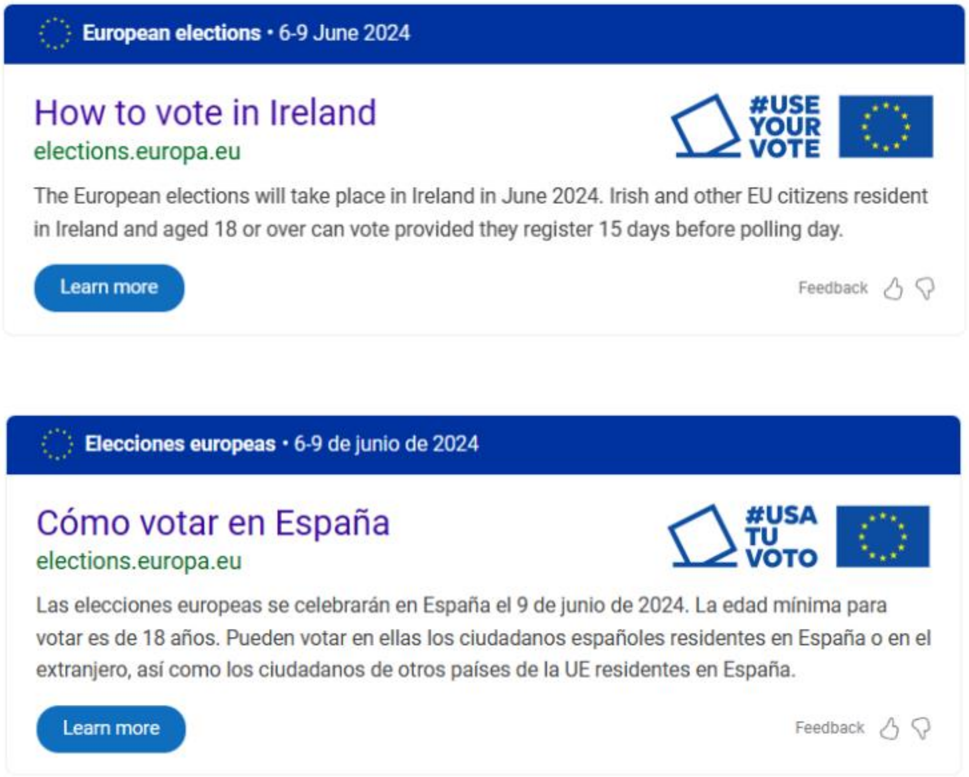
Red Team Testing: Bing Search, Copilot in Bing and Image Creator from Bing and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system. In anticipation of major elections occurring in 2024, in Fall 2023 Microsoft also entered a research engagement with a third party institution to conduct external tests of Image Creator from Bing to assess the risk that it might generate images containing information that could be used to support election-related disinformation campaigns and made product improvements based on the results of the assessment. Microsoft has also worked with reputable third parties to test election-related responses in Copilot in Bing.

Continuous Improvement of Ranking Algorithms: Bing invests significant time and resources into ensuring its crawlers and algorithms prioritize high quality content to avoid inadvertently returning low quality harmful materials to users who have not expressed a clear interest in finding it. Bing works to ensure its interventions are effective in any language and region in which Bing offers the service. Bing regularly measures the efficacy of its ranking algorithms using metrics and makes changes as needed, including regularly reviewed and updated performance targets, and through ingestion of user and stakeholder feedback, including Microsoft policy teams, regulators, and civil society, to continue to iterate and improve on trust and safety in the Bing service.

Defensive Search Ranking with Authority Signal Boost: Bing leverages trusted intelligence sources to inform defensive search mitigations such as giving an extra authority signal boost in search results ranking for queries related to emerging information manipulation terms such as fact checked narratives, hashtags where Bing expects users may be at heightened risk of active information manipulation or data voids. Bing monitors election-related information manipulation themes from reliable external sources such as NewsGuard, GDI, and European Digital Media Observatory (EDMO) covering local intelligence from various EU countries.

Special Authoritative Elections Answer in Search: In March 2024, in advance of EU Parliament elections, Bing launched special informational and Answer segments for the European elections to show high authority information related to the election for users in the EU., localized by language and member state. Bing Search directs users to official sources related to the European elections, such as https://elections.europa.eu/, and other high authority sources. Bing also swiftly built an authoritative "How to Vote" answer to government election resource ahead of the France 'snap' parliamentary election.

Figure 17: Direction to authoritative official election information for EU Parliament elections screenshot



European elections • 6-9 June 2024

How to vote in Ireland

elections.europa.eu

The European elections will take place in Ireland in June 2024. Irish and other EU citizens resident in Ireland and aged 18 or over can vote provided they register 15 days before polling day.

Elecciones europeas • 6-9 de junio de 2024

Cómo votar en España

elections.europa.eu

Las elecciones europeas se celebrarán en España el 9 de junio de 2024. La edad mínima para votar es de 18 años. Pueden votar en ellas los ciudadanos españoles residentes en España o en el extranjero, así como los ciudadanos de otros países de la UE residentes en España.

Trustworthiness Signals for Users: 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. For example, Microsoft partners with NewsGuard to help users evaluate the quality of the news they encounter online. NewsGuard has created trust ratings for 7,500+ news and information sites, which are compiled into a "Nutrition Label" and corresponding Red/Green Reliability Rating to help users understand the reliability of news sources. Within the EU, NewsGuard is currently available in France, Germany, and Italy with plans for future expansion. For users with the NewsGuard plug-in, Bing Search results (including Copilot in Bing) include NewsGuard Reliability ratings that lead to a pop-up screen with more site information.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has created special mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of certain harms and misuse of generative AI features. Classifiers classify text to flag different types of potentially harmful content in search queries, chat prompts, or generated responses. Microsoft uses AI-based classifiers and content filters, which apply to all search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by new generative AI features such as Copilot in Bing. Flags lead to potential mitigations, such as restricting return generated content responses to the user, diverting the user to a different topic, or redirecting the user to traditional search. Metaprompting involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional filtering and classifiers to prevent Copilot in Bing chat responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors. In the leadup to EU Parliament elections, Copilot in Bing developed a novel elections classifier intended to restrict Copilot in Bing from responding to certain election-related queries out of an abundance of caution. This classifier has been a focus of continued improvement, including focused testing in key EU languages used by Bing users in EU.

Leveraging Local Election Data: Copilot in Bing and Image Creator from Bing utilize blocklists to restrict generation of images or certain types of content concerning politicians in upcoming elections as an additional mitigation. Through Microsoft's Democracy Forward team, these services have integrated information on political parties, candidates, and elections from local election authorities (including in the EU) or high authority third party sources to inform defensive interventions and election-related product mitigations.

Content Provenance: Bing is also working to effectively label generated output so that it cannot be used to misinform on other platforms, such as including watermarks on generated images and through participation cross-industry efforts to improve identification techniques like the Coalition for Content Provenance and Authenticity (C2PA). Microsoft and Bing participate in the Partnership on AI ("PAI"), a non-profit that works to identify possible countermeasures against deepfakes. Microsoft and Bing contributed to the development of the Responsible Practices for Synthetic Media guidelines. Microsoft is a contributing member to the NIST AI Safety Institute Consortium (AISIC), working to guidelines related to provenance and watermarking tools and practices that enable the identification of AI-generated or -modified content.

New Feature Launch: Bing relies on Microsoft's extensive cross-company compliance infrastructure to proactively review new products and features to ensure they meet Microsoft's policy commitments in key areas such as privacy, accessibility, digital safety, and RAI.

The Tech Accord to Combat Deceptive Use of AI in 2024 Elections: Microsoft has worked with a number of other leading technology companies to create the Tech Accord to Combat Deceptive AI in the 2024 Elections. The Tech Accord's commitments are designed to make it more difficult for bad actors to use legitimate tools to create politically related deepfakes and other deceptive AI elections content, while simultaneously simplifying the process for users to identify authentic content. The Tech Accord calls on companies that generate AI content and those that distribute it to strengthen the safety architecture of their AI services by assessing risks and enhancing controls to help prevent abuse. Microsoft is one of the founding members of the Tech Accord, and Bing, responsible AI, and generative AI teams have worked to implement Tech Accord pillars into product safety systems.

Microsoft-wide Election Support. An important component of the Democracy Forward team's strategy is to establish lines of communication between election authorities in the EU member states and to identify risks to electoral processes, including possible foreign information operations targeting elections.

While these broader engagements led by Democracy Forward are not specific to Bing (and are part of a whole-of-company approach), they can support and inform risk mitigation on the Bing platform. Recent measures during the reporting period include:

- **Election Communications Hub:** Microsoft created and provided access to a new "Election Communications Hub" to support democratic governments & political parties around the world as they build secure and resilient election processes. This hub has provided election authorities access to rapidly report any issues or concerns to Microsoft security and support teams in the days and weeks leading up to their election, allowing them to reach out and get swift support if they run into major security challenges. Working with authorities and parties in the EU, Microsoft opened a Communications Hub for the EU Parliamentary elections and Member State elections during 2024.
- **Campaign Success Team:** Microsoft helps political campaigns navigate cybersecurity challenges and the new world of AI by deploying a new "Campaign Success Team." This team advises and supports campaigns as they navigate the world of AI, combat the spread of cyber-influence campaigns, and protect the authenticity of their own content and images.
- **Specialized Training:** Microsoft invests in political group awareness campaigns and trainings, such as Deceptive AI and Elections Trainings for political parties and campaigns across the EU to learn how to identify the deceptive use of AI and misinformation in elections and report it with Microsoft's Deceptive AI Reporting Tool (discussed below).
- **Content Credentials as a Service:** Microsoft also launched a broader Content Credentials as a Service ("Content Integrity") to enable political candidates around the world to digitally sign and authenticate media using the C2PA digital credentials. Though this service is Microsoft wide it helps reduce promote content provenance with political candidates.
- **Deceptive AI Election Reporting tool.** During the Reporting Period, Microsoft launched a new web page (available here) where political candidates can report a concern about a deepfake of the candidate on Microsoft services. This empowers political candidates around the world to aid with the detection of harmful deepfakes. This web page is available, localized, throughout the EU. Bing Search, Copilot in Bing, and Image Creator from Bing did not receive any valid deepfake reports from candidates in connection with the June 2024 EU Parliament Election as of this report.
- **Cyber-security and cyber-influence:**
    - **Cyberinfluence Reporting:** Microsoft identifies, tracks, and reports on potential information operations and foreign malign influence operations working with partners such as Spanish news agency EFE and through the MTAC. MTAC's first report, "Protecting Election 2024 from Foreign Malign Influence" was released November 2023, providing a baseline for the following election season, including reflections on previous election influence efforts.
    - **Cyberattack Defense:** Though threats to democracy exist, 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.
    - **AccountGuard:** Microsoft AccountGuard is a threat detection and notification service available at no additional cost to institutions that underpin democracy, including political parties and campaigns, think tanks, human rights organizations, nonprofits, and journalists.

- **Security Guidance:** Microsoft works closely with its elections-related customers to provide security guidance and tools to improve their cyber-resilience and protect the integrity of the electoral process.
- **IFES and NDI Cybersecurity:** Microsoft is supporting IFES to strengthen the cybersecurity practices of investigative journalists who are reporting on abuse of state resources in elections. Microsoft is also partnering with the National Democratic Institute (NDI) to strengthen the cybersecurity support infrastructure for political parties and campaigns internationally

Product Governance

Terms & Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft Services Agreement Code of Conduct broadly prohibits using Microsoft services for misleading or deceptive purposes. In addition to the Microsoft Services Agreement, Users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with deceptive content, including disinformation).

Microsoft Adverting restriction on political ads: Microsoft does not allow political advertising within the MS Advertising ecosystem, which supports advertising on Bing. Microsoft Advertising policies 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 and ballot measures also are barred. Microsoft Advertising's policies also prohibit certain types of advertisements that might be considered issue based. Microsoft Advertising is a signatory to the COPD and more details on how its commitments is contained in Microsoft's most recent COPD report Bing has partnered with Microsoft Research and third-party research organizations to contribute to novel research and internal studies concerning safe design practices, RAI, and disinformation. In preparation for global elections, Microsoft Research recently conducted internal research concerning information integrity and elections in the age of generative AI.

Third-Party Engagement: Bing has partnered with academic institutions like Princeton University, and external organizations like NewsGuard and GDI to further improve its understanding and management of information operations. Bing and Microsoft representatives also participated in EU elections tabletop workshops organized by the Commission dedicated to assessing risk scenarios concerning elections in the European Union.

Transparency Reporting: As part of its commitments under COPD, Bing provides detailed reporting on efforts to mitigate misinformation risks concerning EU elections as well as broader misinformation mitigations and data points. In addition, Microsoft provides transparency on Government Requests for Content Removal (including on Bing). This transparency is in furtherance of Microsoft's commitments to free expression and supports open civic discourse.

Product Enforcement

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing's content removal practices for the search index are largely reactive to reports and involve human review. Bing's support teams are provided with training and oversight to ensure removal practices align with Bing's principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing's generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing's product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing's operations.

Incident Response: Bing maintains an incident response process for cross-functional teams to prioritize high-risk incidents and track the investigation, fixes, and post-incident analysis. Internal escalation processes are set up to ensure urgent cases– including sensitive issues related to elections or election-related content -- are addressed expediently with high priority. Bing also implemented a specialized intake and operations process under the COPD Elections Working Group Rapid Response System and coordinates with Democracy Forward Election Hubs on incidents.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing.

Product Improvement

Monitoring and Testing: Bing monitors metrics, incidents, and potential risks in the product via red teaming, social listening, and other intake channels.

During the Reporting Period, Bing has improved its monitoring capabilities setting up specific measurement on election-related query sets:

- Bing monitors **DDR** on an election query set to measure the presence of content that goes against its search policies on different features of Bing Search (including Web, Image, and Video).
- Bing measures the percentage of responses that are **grounded in web results** following Bing's authority-driven ranking in Copilot in Bing. In sample-based testing, Bing has found that 90%+ of Copilot in Bing responses that were evaluated on a set of elections queries were grounded in web results following Bing's authority-driven ranking.

- **Freshness**: Bing measures if up-to-date results are provided and used in grounding responses with factual information on election query set. Given the rapid pace at which elections and campaigns develop, measuring freshness is critical for this risk.
- In addition to metrics, Bing has tested a high number of prompts related to key global elections in 2024 in multiple languages as part of red teaming and applied fixes as needed.
- Bing monitors jailbreak attempts of Copilot in Bing. Bing has caught more than 6,000 jailbreak attempts in the imminent lead up (June 2-June 8, 2024) to the EU Parliamentary election.
- Bing also uses social listening to evaluate online conversation around potential responsible AI risks regarding election and political topics and enable mitigation actions.

Threat Intelligence: Microsoft works to identify and track nation-state information operations targeting democracies across the world and works with trusted third-party partners, including NewsGuard, Global Democracy Index, and EFE, to provide early indicators of narratives, hashtags, or information operations that can be leveraged to inform early detection and defensive search strategies.

Post-Election Reviews: Bing undertakes post-election reviews, as appropriate, to evaluate product and mitigation performance, reflect on challenges and learnings, and identify potential areas for improvement. These reviews occur both in product review settings and in broader cross-functional teams dedicated to elections at Microsoft.

- Bing and Microsoft representatives have also participated in the multi-stakeholder European Commission Elections “tabletop” exercise ahead of the EU Parliamentary Election, related post-election roundtable session, and meetings with Coimisium na Mean (CnaM) to share practices & reflections.

EU Regulatory Engagement: Bing participated in pre-election discussions and review with CnaM and the European Commission as well as pre-election tabletop exercises prior to the EU Parliament Elections in June 2024 to discuss election-related concerns, highlight Bing election plans, and discuss general election response.

External Collaboration: Both Bing and specialized cross-Microsoft teams that support broadly across the company regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Product Transparency

Research Support: Bing and Microsoft, including through its research arm Microsoft Research, have collaborated with and third-party research organizations to contribute to novel research and internal studies concerning safe design practices, responsible AI, and disinformation. In preparation for global elections, Microsoft Research recently conducted internal research concerning information integrity and elections in the age of generative AI. Bing is evaluating additional research data sharing opportunities pertaining to elections.

Public Health

Risks related to Public Health include exposure to content containing legal but harmful substances and health misinformation. Content related to illegal substances is addressed in the Illegal Content section above. Bing applies its multi-pronged layered safety mitigations to content that poses risks to public health. Bing's ranking algorithms and generative AI safety mitigation systems are designed to protect users from harmful web content and prevent harmful or misleading generative AI outputs. Bing's threat monitoring and incident response processes are designed to continuously monitor evolving trends and place rapid interventions where needed. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Public Health, the Bing Risk Assessment team has assessed the Residual Risk related to Public Health on the service to be Low.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Public Health	Not Likely	Critical	Moderate	Managed	Low

Risk Definition

Risk Area	Public Health
Risk Definition	Risk that content or activities with actual or foreseeable negative effects on the protection of public health occur on the service.
Theoretical Risk Manifestations absent Mitigations	• Risk that websites indexed by Bing Shopping include listings for counterfeit, gray market, or unlicensed pharmaceutical products, medical devices, supplements, or other products that could be detrimental to users' health • Risk that Advertisements on Bing are used to promote legal but harmful substances such as tobacco • Risk that Search results include information about medical conditions, treatments, vaccines, or homeopathic remedies that are inaccurate, unsafe, or otherwise not appropriate or advisable based on an individual's condition • Risk that Search results include low quality information or related to health conditions or medical advice • Risk that bad actors intentionally manipulate search results to increase the visibility of health misinformation • Risk that Copilot in Bing responses include health misinformation or link to sites that contain low authority or misleading health information • Risk that Image Creator from Bing is misused to generate misleading imagery related to public health issues

Risk Analysis

The **Inherent Probability** of systemic risks related to Public Health stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Not Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- A review of internal metrics and social listening data indicates a “Not Likely” level of probability relative to other potential systemic risks on Bing.
- User intent is generally required for users to access potentially harmful content on Bing. Due to the nature of Bing as a search engine, Bing’s goal is to provide fair, balanced, and comprehensive content while respecting user intent. When a user expresses a clear intent to access specific information, Bing provides relevant results even if they are less credible, while working to ensure that users are not misled by such search results. Absent a clear intent to access specific content, Bing assumes an intent to find high authority results. By design, the likelihood of users being exposed to potentially harmful content on Bing without demonstrated intent is limited.
- As Bing generally does not offer features for users to post or share their own content, or engage with other users on Bing, the probability of content "going viral" and thereby leading to systemic impacts is reduced, which has a tempering impact on the probability score.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** for content and activities that negatively impact Public Health on Bing is rated as “Critical,” primarily due to the gravity of risks within this category, considering the potential for significant and irremediable harm to wellbeing, economic, and societal systems at the individual and up to global level. Some risks within this category such as health misinformation and exposure to harmful substances have potentially irremediable consequences.

With the rating for Inherent Probability as “Not Likely” and Inherent Severity as “Critical,” the **Inherent Risk** associated with Public Health on the Bing service is “Moderate.”

Risk Mitigation

Bing has implemented a robust set of mitigations to address risks related to Public Health, including those described in Bing’s approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and those summarized below. The mitigations Bing has implemented relative to Public Health follow best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts Bing has applied to the Public Health risk area is assessed as “Managed,” which brings the **Residual Risk** rating for down to “Low.”

While not a proxy for Residual Risk, Bing has measured the DDR for Public Health on Bing Search as an average of .65% from April to June 2024. This means that Bing estimates .65% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Public Health.

Nevertheless, Bing continues to invest in developing and enhancing strategies to further mitigate and manage risks related to Public Health. The key mitigations currently implemented are described below.

Product Development

Algorithmic Prioritization of High Authority Content: To provide users with the most relevant and highest authority content in search queries, Bing invests significant time and resources into ranking and relevance systems. Bing regularly reviews product quality issues, recurring trends, and emerging risks to ensure its algorithms are sufficient without subjecting the service to unnecessary censorship. In most cases, the user will be able to find the content in standard search results regardless of content being removed from generative AI features. Bing regularly measures the efficacy of its ranking algorithms using metrics and makes changes as needed, including regularly reviewed and updated performance targets, and through ingestion of user and stakeholder feedback, including from Microsoft policy teams, regulators, and civil society, to continue to iterate and improve on trust and safety in the Bing service.

RAI Program: Microsoft processes, programs, or tools utilizing AI, including Copilot in Bing, Image Creator from Bing, and other Bing search features, must adhere to Microsoft's RAI Standard and undertake RAI review to help ensure responsible use of AI-influenced algorithms and processes for any new product features prior to launch.

Grounding GenAI in High-Ranking Search Results: Copilot in Bing textual responses are generally grounded in web search results. "Grounding" refers to the process of providing data to an LLM to improve its ability to provide accurate, relevant, and updated outputs. This means that responses to user prompts are centered on high authority content from the web (except for creative use), and Copilot provides links to websites so that users can learn more and evaluate the credibility and information presented by reviewing the source material.

Red Team Testing: Bing Search, Copilot in Bing and Image Creator from Bing and their features are required to conduct pre-launch and ongoing testing. For example, before launching Copilot in Bing (then known as Bing Chat), Microsoft conducted extensive "red team" testing. A multidisciplinary team of experts conducted numerous rounds of testing to evaluate how well the system responded when pressed to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations. Post-release, Copilot in Bing experiences are integrated into Microsoft engineering organizations' existing production measurement and testing infrastructure. Red team testers from different regions and backgrounds continuously and systematically attempt to compromise the system, and their findings are used to expand the datasets that Microsoft uses for improving the system.

Dedicated Expert Teams: Bing has a dedicated team that is accountable for implementing algorithmic interventions and metrics monitoring and dedicated to identifying and remedying high impact issues in search results, such as misinformation, public health concerns, self-harm materials, and other problematic content that could negatively impact public health by deploying targeted algorithmic interventions. This team identifies high-risk areas based on emerging threats and misinformation narratives to investigate and remedy circumstances where algorithmic results are inconsistent with Bing's ranking principles and goals. Examples of defensive search algorithmic interventions can include QC boosts for authoritative websites, demotions of a low authority websites, restricting search suggestions to avoid directing users to potentially problematic queries and engaging in manual interventions for specific reported issues or in broader areas more prone to misinformation or disinformation (e.g., vaccines or COVID-19). This team relies upon red team testing, external threat intelligence sources, social listening systems, consistent monitoring of manipulation trends, and other mechanisms to help maintain the integrity of Bing algorithms and search results. These measures are deployed and localized in any language and region where Bing is offered.

Image Creator from Bing Blocklists: Image Creator from Bing uses blocklists to block the generation of content based on user inputs that contain certain names or terms and these user inputs are placed in "hard block list" or "soft block list" based on their classification.

New Feature Launch: Bing relies on Microsoft's extensive cross-company compliance infrastructure to proactively review new products and features to ensure they meet Microsoft's policy commitments in key areas such as privacy, accessibility, digital safety, and RAI. For example, privacy reviews are an essential part of any new feature review process; implementing recommended privacy mitigations following such reviews is a requirement for launch. Microsoft has a robust privacy and security infrastructure, consisting of privacy managers who are trained in Microsoft privacy standards and relevant laws, and have access to centralized privacy specialist teams and legal support for complex or novel issues.

Product Governance

Terms & Conditions: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in any member state languages. The Microsoft Services Agreement Code of Conduct broadly prohibits using Microsoft services for misleading or deceptive purposes. In addition to the Microsoft Services Agreement, Users of Copilot services are subject to Copilot AI Experiences Terms; users of Image Creator from Bing services are subject to Image Creator Terms. Users may be banned from the service for attempting to use the service in ways that violate these terms, including users who enter prompts that are designed to bypass Copilot's safety systems or create content that violates the Code of Conduct (which further prohibit use of the service in connection with deceptive content, including counterfeit drug listings, and health misinformation).

Microsoft Advertising Policies: Microsoft Advertising, which powers ads on Bing, has clear and regularly enforced content policies that prevent advertising of harmful materials, including illegal products, illegal drugs and drug paraphernalia, tobacco products, hazardous materials, harmful/misleading content, and other content that could threaten physical, mental, or personal safety.

Policy Violating Content Monitoring: Bing proactively uses hash-matching technologies (including PhotoDNA and MD5) to detect matches to known CSEAI, to avoid it from appearing in the index using PhotoDNA and via threat intelligence provided by third-party expert partners. Bing also uses these technologies to block CSEAI from being uploaded for use in Bing's visual search feature to prevent users from seeking harmful materials and support public health.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of QC score, which is one of Bing's main ranking parameters. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

Additional Algorithmic Interventions:: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or "defensive search interventions," to counteract these threats, in accordance with its search principles.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing's content removal practices for the search index are largely reactive to reports and involve human review. Bing's support teams are provided with training and oversight to ensure removal practices align with Bing's principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of harm and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing's generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing's product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing's operations.

Product Improvement

Internal Metrics: Bing maintains a set of metrics to track, monitor and review the efficacy of its interventions on an ongoing basis to ensure that they are performing as expected and not inadvertently introducing additional bias or causing other harm.

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of negative impacts to Public Health on Bing's services.

External Collaboration: Both Bing and specialized cross-Microsoft teams that support broadly across the company regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, responsible AI and AI-specific risks, minors, and technology, to ensure that Bing internal policies, practices, and standards are addressing key concerns of these stakeholders. These engagements can inform processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. External engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or other special information "hubs" related to public health issues. For example, during the peak of the global COVID-19 pandemic, Bing offered special answers and a COVID-19 information hub for users containing news and data from high authority sources to provide reliable sources of information to users.

Content Provenance & Watermarks: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

AI Disclosures: Copilot in Bing provides in-service disclosures and reminders to users that AI can make mistakes and to double check the veracity of information generated by AI. In addition, the Copilot in Bing FAQs, help pages, and other public facing information sources (such as blog posts and marketing materials) help educate users on the nature of AI-driven search experiences and the uses, safeguards, and limitations of this emerging technology, regularly reminding users of the potential for mistakes and risks of over-reliance.

Transparency Reports: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in EU member state languages. Where content is removed from search, Bing is transparent by notifying users through a notice on the search engine results page and publishing regular transparency reports (available on Reports Hub | Microsoft CSR). Microsoft's Reports Hub is a publicly available source where users can access various transparency reports in areas where Bing reports on various practices. These key areas include: RAI, content removal requests (such as copyright or digital safety), privacy (such as the "Right to be Forgotten") and security (such as Government Requests) The Reports Hub contains additional transparency reports such as jurisdictional, community and privacy and security transparency reports that users can easily access. As a result of the conduct of the Systemic Risk Assessment, each year Bing identifies specific areas for focused enhancements in the coming year.

EU Code of Practice on Disinformation: Bing is a signatory to the COPD and provides robust biannual reporting pursuant to the code on systems and policies in place to prevent misinformation on Bing. Until recently, biannual COPD reports included detailed information specifically on measures concerning the COVID-19 pandemic crisis, providing heightened transparency on how Bing counters misinformation and public health risks. Bing is also a joint coordinator in the generative AI working group under the COPD.

Freedom of Expression and Information

One of the Trustworthy Search principles that guides Bing is to promote free and open access to information within the bounds of the law, with respect for local law and other fundamental rights. This is reflected in its careful efforts to not unduly restrict important interests such as freedom of expression, open access to information, and media pluralism. Given that search engines are the primary way users find information online and conduct research relevant to their daily and professional lives, over-moderation of content in search could have a significant negative impact on the right to access information and freedom of expression. Bing must carefully balance these competing fundamental rights and interests as it works to ensure that its algorithms return provide high quality, relevant content in response to the user's queries without unduly limiting their ability to access information or express themselves.

Bing endeavors to provide open access to as much of the web as possible, but in limited cases may undertake certain interventions (such as removal of a website or downranking) such as where the content violates local law, or Microsoft's policies. When limiting access to content, Bing strives to ensure its actions are narrowly tailored so fundamental rights are not unduly restricted. In some cases, different features may require different interventions based on functionality and user expectations. For example, in some cases, rather than removing or blocking content, Bing may inform users of certain risks through public service announcements or warnings or provide users with options for tailoring their content. Bing

Following Bing's principles, Bing's safety systems are balanced against the rights to freedom of expression and information. Bing provides transparency around Bing's safety practices, content moderation processes and content removal data. For generative AI features, although Microsoft may take more pre-emptive action to limit generation of certain types of harmful content, it carefully measures and monitors overblocking rates (i.e. rates at which responses are restricted) to drive improvements. Considering the Inherent Probability, Inherent Severity, and the Maturity of Mitigations relevant to Freedom of Expression and Information, the Bing Risk Assessment team has assessed the Residual Risk related to Freedom of Expression and Information on the service to be Low.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Freedom of Expression and Information	Not Likely	High	Moderate	Optimized	Low

Risk Definition

Risk Area	Freedom of Expression and Information
Risk Definition	Risk that content or activities with actual or foreseeable negative effects on the fundamental rights to freedom of expression and information, including the freedom and pluralism of the media occur on the service.
Theoretical Risk Manifestations absent Mitigations	• Risk that Copilot in Bing and Image Creator from Bing disproportionately block certain prompts or generated responses • Risk that Search's ranking processes disproportionately favor or demote certain types of content or viewpoints • Risk that Search over moderates the search index, removing or downranking URLs limiting access to information • Risk that the mitigations in Copilot in Bing and Image Creator from Bing (such as blocklists, filtering, classifiers, metaprompts and other safeguards) disproportionately restrict certain types of content users seek • Risk that Image Creator from Bing or Copilot in Bing disproportionately generates images that align with particular cultural standards or viewpoints, resulting in less diverse representations of the population

Risk Analysis

The **inherent probability** of systemic risks related to Freedom of Expression and Information stemming from the use, misuse, or functioning of Bing's products and services absent sufficient mitigations is assessed as "Not Likely." The Bing Risk Assessment team considered the following factors in assigning this score:

- Absent mitigations to ensure that Bing, in the application of its controls and safeguards, does not overly restrict freedom of expression or information, the likelihood of this occurrence is Remote as Bing generally does not remove third-party website from the Search index, absent legal or policy reasons. The likelihood of Bing overly restricting freedom of expression or information on GenAI features is greater, but this is also balanced by the fact that users can continue to search for information using Bing web search (or other search mediums), even where responses or outputs are restricted in generative AI features. Additionally, a review of internal metrics and social listening data indicates a "Not Likely" level of probability relative to other potential systemic risks on Bing.
- A lower percentage of Bing search users make use of generative AI and ancillary search features. Roughly 15% of Bing Search users in the EU also use Copilot in Bing and around 2% use Image Creator from Bing. Roughly 10% of Bing Search users in the EU use Maps, 4% use Shopping, 2% use News, 1% use Travel, and .25% use Real Estate. Therefore, while the inherent probability and ways that users may be exposed to potentially harmful content vary between core Search, generative AI features, and ancillary search features, the probability score is weighted toward occurrence on Bing Search.

The **Inherent Severity** for content and activities that negatively impact Freedom of Expression and Information on Bing is rated as "High" primarily due to the gravity of risks within this category, considering the potential for significant harm to individuals' fundamental rights as well as broader societal, media, and political systems that built upon the rights to free expression, access to information, and media plurality. Consequences related to the negative impact on Freedom of Expression and

Information can be irremediable at scale. But the impact of many risks within this category is not considered significant.

With the rating for Inherent Probability as “Not Likely” and Inherent Severity as “High,” the **Inherent Risk** associated with Freedom of Expression and Information on the Bing service is “Moderate.”

Risk Mitigation

Bing has implemented an expansive set of mitigations to address risks related to Freedom of Expression and Information, including those described in Bing’s approach to risk mitigation and in the Catalog of Mitigations by Industry Best Practices and those summarized below. The mitigations Bing has implemented follow best practices promoting freedom of expression. As such, the **maturity of mitigation** efforts Bing has applied to the Freedom of Expression and Information risk area is assessed as “Optimized,” which brings the **Residual Risk** rating for down to “Low.”

While not a proxy for Residual Risk, Bing has measured the DDR for Freedom of Expression and Information on Bing Search as an average of .83% from April to June 2024. This means that Bing estimates .83% of search traffic may include harmful content leakage, where users are unexpectedly exposed to content that may have negative effects related to Freedom of Expression and Information.

Nevertheless, the rights to free expression and access to information remain fundamental to Bing’s approach to online safety. Bing continues to invest in innovating and enhancing strategies to further mitigate and manage risks related to Freedom of Expression and Information, particularly with respect to generative AI features that offer new ways for users to approach research, learning, and search. The key mitigations currently implemented are described below.

Product Development

Trustworthy Search Principles: As a search engine, the fundamental rights to access and seek information and freedom of expression are core to Bing’s Trustworthy Search principles. Bing limits removal of content to narrow scenarios to avoid unduly impacting access to information. Bing provides training and oversight to its content review teams to ensure consistent application of policies, including reviews of decisions, and provides escalation paths to local legal experts as needed. For government demands, Bing employs additional safeguards to ensure any actions taken are narrow, specific, submitted in writing, and based on valid legal orders. Bing has automated safeguards in place to trigger additional reviews as needed. Where content is removed from search, Bing is transparent by notifying users through a notice on the search engine results page and publishing regular transparency reports (available on Reports Hub | Microsoft CSR).

Bing News Safeguards: Similarly in Bing News, Bing supports freedom of information and media pluralism, by providing users with a variety of high authority news resources and has established safeguards to avoid inadvertently presenting users with low authority content, including: 1) requiring Bing news search results to meet specific Bing News Publisher Guidelines, 2) working with in-market teams to identify high trust news sources for content that is recommended to users, and 3) enabling the NewsGuard plugin to help users identify the reliability of the sources they are accessing.

Generative AI Safeguards: In addressing the complexities of sensitive topics, Microsoft's generative AI features are designed to direct users to authoritative sources and to decline to respond where the risks to user safety are greater than the risks to freedom of information. However, Microsoft continuously measures not just “leakage,” where defects are present in generative AI content, but also “overblocking,” where sample-based review of prompts and responses determines that Copilot in Bing or Image Creator from Bing are unnecessarily blocking responses. Microsoft tracks these two metrics alongside one another and continuously adjusts and refines classifiers and metaprompts to strike the right careful balance.

Global Network Initiative Audits: Microsoft also is a member of the GNI, a multistakeholder collaboration to protect freedom of expression and privacy in tech and submits to a regular audit by GNI to ensure Bing has sufficient protections in place to uphold free expression. Bing is biannually audited for its commitments to human rights, including free expression, as part of its membership in the GNI. Microsoft also is a member of the GNI, a multistakeholder collaboration to protect freedom of expression and privacy in tech and submits to a regular audit by GNI to ensure Bing has sufficient protections in place to uphold free expression.

New Feature Launch: Bing relies on Microsoft's extensive cross-company compliance infrastructure to proactively review new products and features to ensure they meet Microsoft's policy commitments in key areas such as privacy, accessibility, digital safety, and RAI Product Governance

Search Principles: Central to Bing's strategy is a set of search principles that respects the freedom of expression and information while balancing it with other key interests and the risks of harm occurring on its service. Bing continuously engages in internal and external discussions to evaluate and refine the principles.

Content Removal Policies: Developing policies and employing a cautious approach toward content removal, Bing maintains minimal, justified intervention, consistently informing users about removed content through notices and transparency reports. Bing emphasizes presenting varied perspectives in search results, particularly valuing media pluralism within Bing News by enforcing strict publisher guidelines and promoting reliable news sources.

Product Governance

Central to Bing's strategy is a set of search principles that respects the freedom of expression and information while balancing it with other key interests and the risks of harms occurring on its service. Bing continuously engages in internal and external discussions to evaluate and refine the principles.

Product Enforcement

Search Algorithms: Search algorithms and recommender systems are the most fundamental part of Bing's risk mitigation approach. Bing designs its ranking and recommendation algorithms to align with core product principles that prioritize high quality, relevant content, and to ensure that users are not offended, harmed, or misled by problematic material in search results. Bing builds and maintains systems to consume external signals to identify the authoritativeness of websites and use the authority as a part of QC score, which is one of Bing's main ranking parameters. Bing rigorously measures and monitors metrics to identify the implementation gaps and inform the product strategy.

Additional Algorithmic Interventions: Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that seek to undermine the efficacy of these algorithms, including where there are data voids that are prone for search engine exploitation. Bing deploys additional algorithmic interventions, or "defensive search interventions," to counteract these threats, in accordance with its search principles.

Content Moderation: In limited cases, Bing removes content from the index for legal or policy reasons. While Bing employs some automated content detection that identifies with a high degree of accuracy crawled content that should be excluded from the index, such as spam and child sexual exploitation and abuse imagery, in most cases automated content detection is not feasible to use for content removal decisions, as most content removal decisions are highly context dependent. As a result, Bing's content removal practices for the search index are largely reactive to reports and involve human review. Bing's support teams are provided with training and oversight in order to ensure removal practices align with Bing's principles and legal obligations, and have escalation paths for difficult issues, including consultation with local legal experts where needed.

Monitoring and Enforcement: Bing monitors for violations of its Webmaster Guidelines (e.g., improper attempts to manipulate search algorithms via keyword stuffing or other prohibited practices) and terms of use, and actions violations consistent with its policies and procedures. Bing also maintains a social listening pipeline where insights and user feedback on Bing's generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing's product teams and product engineering leadership in a daily report. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing's operations.

Incident Response: In addition to algorithmic ranking intervention and reactive content removal, Bing has an internal escalation process set up to help ensure that urgent cases where users rights are impacted are being investigated and fixes or improvements are made where applicable with high priority.

User Reporting and Feedback: At the bottom of each page on Bing, users can find a link entitled "Feedback." Users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal policies to ensure it is appropriately routed and actioned. In addition, where users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through the Report a Concern tool, accessible through the Feedback link available on each page of Bing.

Classifiers, Metaprompting, and Filtering Interventions: Microsoft has implemented mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of harm and misuse specific to generative AI features. **Classifiers** classify text, image, and video to flag different types of potentially harmful content in search queries, chat prompts, or generated responses or image output. Microsoft uses AI-based classifiers and content filters, which apply to search results and relevant features; it also designed additional prompt classifiers and content filters specifically to address possible harms raised by generative AI features such as Copilot in Bing and Image Creator from Bing. Flags lead to potential mitigations, such as restricting returning generated responses/images to the user, diverting the user to a different topic, or redirecting the user to traditional search. **Metaprompting** involves giving instructions to the AI model to guide its behavior, including so that the system behaves in accordance with Microsoft's AI Principles and user expectations. Microsoft has also implemented additional **filtering** and classifiers to prevent Copilot in Bing responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

News Content Review: Bing scans news content by standard Bing classifiers to flag content for action, and to filter out lower quality sources. Bing's nuanced approach ensures that, except in cases where content is illegal or violates another fundamental right such as privacy or human dignity, information remains accessible. For government demands, Bing employs additional safeguards to ensure any actions taken are narrow, specific, submitted in writing and based on valid legal orders.

Operationalization of Policies: The operationalization of policies comes through in Bing's handling of specific scenarios, such as compliance with the EU's "Right to be Forgotten". Frequent training for content review teams ensures they have the knowledge and tools they need to apply policies consistently. By dedicating specialized teams to review and action these requests, Bing balances the right to privacy with the public's right to know.

Algorithmic Interventions: Bing also employs a specific team which is accountable for implementing algorithmic interventions and metrics monitoring and remedy algorithmic issues in high priority subject areas and deploys targeted algorithmic interventions as needed to ensure users can access the information they are seeking.

Feature Safeguards: On Copilot in Bing and Image Creator from Bing, Bing prioritizes user safety and security over freedom of information due to the nature of those products. For example, for prompts related to sensitive topics, Bing often redirects users to Search, rather than allowing an AI-generated response, and measures the performance of its overblocking of responses to ensure that. Bing prioritizes freedom of information and pluralism for search, allowing a greater variety of results and news entities to appear when users search for them, however for the News feature Bing is more restrictive and seeks to recommend users with high quality and authoritative sources. Bing balances this prioritization by 1) working with **in-market teams** to identify high trust news sources for content that is returned in results for users queries and 2) attaching Content Credentials, or embeds content provenance credentials, on images generated by Image Creator from Bing and Copilot in Bing to help users identify the reliability of the sources they are accessing.

Product Improvement

External Collaboration: Collaboration with third parties and adherence to **external evaluations** further reinforce Bing's commitment to their policies. By working with external partners to ensure the quality and reliability of content, Bing strives to maintain a high standard for the information it surfaces. Moreover, participating in initiatives like the **GNI**, which audits Bing, and undergoing independent assessments by organizations like **Ranking Digital Rights**, Bing subjects its practices to scrutiny, ensuring they align with global standards for freedom of expression and information.

Effectiveness Testing: Bing routinely reviews the effectiveness of Bing's safety system through internal metrics as well as social listening channels, where insights and user feedback on Bing's generative AI features are collected from the open Internet, to help ensure that its mitigations are working effectively to reduce the risk of negative impacts to Freedom of Expression and Information on Bing's services.

Product Transparency

Special Answers and Panels: Bing may add special answers, news carousels, or other special information panels derived from high authority sources to help direct users to reliable information. In limited circumstances, Bing may also offer warnings or other special information "hubs" related to public health issues. For example, during the global COVID-19 pandemic, Bing offered special answers and a COVID-19 information hub for users containing news and data from high authority sources to provide reliable sources of information to users.

Content Provenance & Watermark: Bing invests in helping users identify AI-generated content and mitigate the risks of misinformation. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allows verification of AI-generated images even after minor modifications. The Watermark complements Bing's content provenance technology to ensure an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques like the C2PA.

Transparency Reports: Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in EU member state languages. Where content is removed from search, Bing is transparent by notifying users through a notice on the search engine results page and publishing regular transparency reports (available on Reports Hub | Microsoft CSR). Microsoft's Reports Hub is a publicly available source where users can access various transparency reports in areas where Bing reports on various practices. These key areas include: RAI, content removal requests (such as copyright or digital safety), privacy (such as "Right to be Forgotten") and security (such as Government Requests.)The Reports Hub contains additional transparency reports such as jurisdictional, community and privacy and security transparency reports that users can easily access. As a result of the conduct of the Systemic Risk Assessment, each year Bing identifies specific areas for focused enhancements in the coming year.

Additional insights

Changes to Bing's risk profile

The primary changes to the risk ratings in this year's Systemic Risk Assessment are due to the enhanced methodology implemented in conducting the Systemic Risk Assessment and not due to a change in the overall risk profile. Bing engaged external risk assessors consulting firm to enhance the methodology and implementation of the 2024 Systemic Risk Assessment, as described in the Methodology section of this report. While this year's assessment results do not significantly deviate from last year's, some of the Residual Risk Scores have shifted, as visualized in Figure 19, reflecting adjustments in the risk areas and scenarios considered, the incorporation of data into the probability scoring, and a more extensive mapping of the potential presentation of various risks across Bing products and features services absent safeguards to the safeguards, controls, and other mitigation measures Bing has implemented. Moving forward, and subject to feedback from the European Commission or third-party auditors, Bing will leverage continue to use this methodology and these identified risk areas to enable a more consistent evaluation of year-over-year changes.

Figure 19: Changes from 2023 to 2024 Systemic Risk Assessment Scores

Risk Area	Inherent Risk Rating	Residual Risk Rating
Civic Discourse and Electoral Processes	High	Low
Consumer Protection and Fraud	**High +**	Low
Discrimination and Hate	**High +**	Low
Freedom of Expression and Information	**Moderate -**	Low

Human Dignity	**Critical ++**	**Moderate +**
Illegal Content and Activities	**High +**	Low
Mental and Physical Wellbeing	High	**Moderate +**
Private and Family Life	**High +**	Low
Protection of Personal Data	**High +**	Low
Public Health	**Moderate -**	Low
Public Security	**Critical +**	**Moderate +**
Rights and Protection of Minors	High	**Moderate +**

Key: + indicates an increase of a score from last year's assessment by one rating level, ++ indicates an increase by two rating levels, - indicates a decrease of a score by one rating level

Alignment of Investments to Risk

While the enhancements made to Bing's Systemic Risk Assessment methodology and process resulted in increases to some Inherent Risk scores, there was not a corresponding proportionate increase in each Residual Risk score. This is a demonstration of the alignment of Bing's investments in safeguards, controls, and other mitigations to the areas of greater risk on the Bing service.

As described in the Monitoring section of this report, Bing leadership routinely tracks areas of systemic risk in order to adjust policies, enforcement, system design and other mitigations to address areas where risks are increasing or changing.

As an example, the higher level of Inherent Risk rating for Discrimination and Hate and Protection of Personal Data is met with a higher Maturity of Mitigations to retain a Low Residual Risk Score despite an increased Inherent Risk Score.

Emerging Trends and Shifts in Global Risk Posture

Various events around the world can also impact Bing's risk profile, including geopolitical, health, and climate events as well as emerging technologies and trending social movements. As part of the 2024 Systemic Risk Assessment, the Bing Risk Assessment team considered a variety of global events to determine whether each warranted a shift in the Inherent Risk score. Examples of the events considered include:

- An unprecedented degree of elections around the world during the Reporting Period;
- Continued geopolitical unrest and active armed conflicts; and
- Expanded use of generative AI capabilities by society and potentially bad actors seeking to leverage these new technologies for harmful purposes.

The Bing Risk Assessment team did not shift any of the probability scores for this year's Systemic Risk Assessment as a result of consideration of the identified global events as the relevant Inherent Risk scores were determined to already be sufficiently high. For example, while the Bing Risk Assessment team considered the unprecedented quantity of global elections this year in the impact on Systemic Risk, that fact did not subsequently reflect a higher Inherent Risk score for Civic Discourse and Electoral Processes as the Inherent Risk was already High. However, the team did ensure specific mitigations were identified to address presentations of the relevant risks across Bing features.

Shifts in Inherent Risk

The Inherent Risk scores for Consumer Protection and Fraud, Discrimination and Hate, Human Dignity, Illegal Content and Activities, Private and Family Life, Protection of Personal Data, Public Security, and Rights and Protection of Minors increased from last year due to Bing's updated risk assessment methodology for assessing Inherent Risk, including the addition of various risk scenarios not considered in the prior year's risk assessment and the consideration of the probability of those risks occurring on or through Bing products and services absent safeguards based on internal and external data sources.

The scores for Inherent Risk of Public Health decreased from last year due to an adjustment of the likelihood of the occurrence of these harms on the service based on internal and external data.

Shifts in Residual Risk

The scores for Human Dignity, Mental and Physical Wellbeing, Public Security, and Rights and Protection of Minors increased from last year due to an increase in the assessed Inherent Risk of the Risk Area and a Defined or Managed Mitigation Maturity defined by an enhanced mapping of implemented mitigations to the identified risk scenarios.

Progress on enhanced mitigations from Bing's last report

In last Bing's first Systemic Risk Assessment report, Bing outlined seven primary areas for enhancement of safety mitigations during the initial reporting period. Bing has made enhancements in each identified area as described below.

1. Continue to navigate the balance of providing broad access to information with the need to protect users from harmful content and mitigate systemic risks:

- Bing continues to follow its authority principles and conduct algorithmic interventions based on its policies. Bing monitors relevant metrics weekly to measure the effectiveness of such interventions and makes continual improvements to the techniques where gaps are observed.
- For Copilot in Bing, Bing also measures the "Overblocking Rate" to ensure that safety mechanisms in the generative AI feature such as disengagement from the conversation, canned response, filtering suggestions are proportionate, in order to balance user safety with rights to access information.
- Bing also continues to monitor incidents, user reports, and social feedback to ensure that its policies and mitigations address emerging risks and trends.
- Bing has engaged internal and external third-party experts to advise on election and political content.
- Microsoft has organized elections-specific working groups to address issues and risks arising from global elections, including teams dedicated to 2024 Elections and specialized AI & Elections working groups. These working groups incorporate and share information with key, relevant Bing personnel.

2. Continue to expand transparency of Bing's principles, processes, and content moderation activities:

- Bing ensures that its search transparency document is up to date and easy to read, and provides users transparency of Bing's principles, processes, and content moderation activities. How Bing delivers search results - Microsoft Support.

- Microsoft regularly updates the Copilot in Bing FAQs help users understand the generative AI technologies and its safety approaches as well as a range of blog posts and other informational resources. (See Your Everyday AI Companion | Microsoft Copilot)
- Bing improved internal processes regarding terms of use updates.
- Bing is transparent about its policies and actions with respect to user and webmaster content and any EU member state language. Microsoft's Reports Hub is a publicly available source where users can access various transparency reports in areas where Bing reports on various practices. These key areas include RAI, content removal requests (such as copyright or digital safety), privacy (such as the "Right to be Forgotten") and security (such as Government Requests.) The Reports Hub contains additional transparency reports such as jurisdictional, community, privacy, and security transparency reports that users can easily access.

3. Expand mechanisms for gathering and seeking out feedback from users, civil society, regulators, industry partners, and other stakeholders, and create more formal processes for reflecting this feedback, as appropriate, in Bing's product improvement processes.

- Bing has expanded external stakeholder engagement mechanisms in various forms to seek feedback and facilitate discussions on topics related to online harms to inform its safety strategy (see examples below). Bing will continue to actively facilitate such engagements and formalize the internal processes to document the engagements.Examples of engagement include:
    - UN INCB on online dangerous substances trafficking.
    - Lucy Faithfull Foundation consulted on Bing Public Service Announcements for CSEAI risk queries which Bing integrated into user experience improvements.
    - Digital Trust and Safety Partnership-Global Network Initiative's European Rights and Risk Assessment forum across DSA systemic risks

4. Expand Bing's relationships with third-party experts to quickly identify new areas of concern and identify and implement more effective interventions for issues of significant concern, including coverage across EU markets and languages.

As discussed above, Bing has expanded relationships with third-party experts through more frequent and deep-dive engagements to identify emerging risk areas and implementation gaps. For more information, see the External Consultation section in Appendix II.

Representative examples of external engagements undertaken during the Reporting Period include: Global Internet Forum to Counter Terrorism (GIFCT), the WeProtect Global Alliance, the Christchurch Call to Action~~,~~ the World Economic Forum's Global Coalition on Digital Safety, the Organisation for Economic Co-operation and Development (OECD), the IGF, the Global Project Against Hate and Extremism, the European Accessibility Summit, the Partnership on AI, the NIST AI Safety Institute Consortium (AISIC) synthetic media working group, the GDI, Truepic, MSR, Princeton University, Freedom Online Coalition, the Global Network Initiative (GNI), NewsGuard, the Digital Trust & Safety Project, the Advisory Network of the Freedom Online Coalition, the United Nations' International Narcotics Control Board, the Anti-Defamation League, AlgorithmWatch, the Better Internet for Kids Forum, the Family Online Safety Institute, the Tech Accord to Combat Deceptive Use of AI in 2024 Elections, the EU Code of Practice on Disinformation, the European Digital Media Observatory (EDMO) Working Group for the Creation of an Independent Intermediary Body to Support Research on Digital Platforms (IIB), and the EU Council's

"Horizontal Working Party on Enhancing Resilience and Countering Hybrid Threats" ("ERCHT"), among others.

5. Expand Bing's network of relationships with internal and external researchers, including creating more formal processes by which Bing can share data with researchers.

- Bing formalized its Qualified Researcher Program to enable EU researchers to easily request access for publicly accessible Bing data from a dedicated landing page and looks forward to supporting vetted researcher process once it is formally established.
- Microsoft collaborated with researchers from the University of California – Irvine's CERES researching network ("Connecting EdTech Research EcoSystem") to identify how to address the needs of child and teen users of generative AI in search.

6. Further document policies, processes, and responsibilities for Bing's compliance and safety operations to ensure internal accountabilities and adherence to procedures, including improving processes for handoffs between internal teams, and further advance Microsoft's and Bing's compliance functions:

- Bing has improved the internal processes and documentation related to ongoing safety and compliance efforts.
- Bing developed improved audit controls and processes to help measure and map compliance with its obligations under the Digital Services Act
- Bing partnered with internal teams in Democracy Forward, Legal Affairs, the Office of Responsible AI, Digital Safety, Government Affairs and other relevant engineering groups to address elections-related issues and established clear working groups and responsibilities in preparations for key upcoming elections in the EU

7. Take a leadership role in the industry's approach to integrating safety and risk reduction into generative artificial intelligence and emerging technologies:

- Microsoft acted as leader in driving the Tech accord to Combat Deceptive Use of AI in 2024 Elections, which supports broader generative AI safety on Bing and other Microsoft Services
- Bing has engaged with Better Internet for Kids and the Family Online Safety Institute to gather feedback on Bing's AI-driven search services directly from teens in support of improving its child safety measures.
- Microsoft and key members of the Bing Search team are involved in the Partnership on AI, including efforts to identify possible countermeasures against deepfakes through the drafting of proposed Responsible Practices for Synthetic Media.
- Bing has consistently evaluated and improved novel safety mitigations for generative AI features, including through the development and enhancement of metrics and monitoring, classifiers, filtering, metaprompts, and blocking technology.

Looking ahead to the coming year

Although Bing has made investments into digital safety and risk mitigation throughout this reporting period, it looks forward to continuing to build upon and improve its approach to mitigating systemic risks in the Union and adapting to new challenges ahead.

The following areas have been identified for heightened focus for the next reporting period, in addition to standard risk mitigation monitoring and refinement processes. Bing is committed to addressing each of the topics identified below by establishing a working group to 1) further define and evaluate the associated risks, 2) explore enhanced mitigation options, 3) review mitigation options as appropriate with internal and external subject matter experts (as appropriate) 4) develop action plans as needed to address focus areas, and 5) track implementation of the developed action plans.

General Focus Areas:

- Improved mapping of existing risk mitigation and monitoring processes to individualized systemic risks.
- Solidify safety governance across newly created Microsoft AI organization, of which Bing is now a part.
- Increased investment in documentation of relevant policies, processes, practices, and procedures, across systemic risk categories, including product mitigations and measurement.
- Refined user reporting capabilities across Bing enhanced search features.
- Increased engagement with experts to improve understanding of how risk areas manifest or are experienced by Bing users to consider additional or refined mitigation measures.

Consumer Protection and Fraud:

- Continued investment in classifiers and mechanisms for identifying evolving consumer protection risks on the service, including malware.

Human Dignity:

- Continued investment in classifiers and mechanisms for identifying evolving harms on the service, including content related to sexual exploitation.

Mental and Physical Wellbeing:

- Continued investment in classifiers and mechanisms for identifying evolving harms on the service, including topics related to suicide, self-harm, and eating disorders.

Public Security:

- Further enhance crisis and rapid response protocols across Bing ecosystem, including ancillary search features.

Rights and Protection of Minors:

- Expanded protections across enhanced search features to further address safeguards for minor users.

Risk assessment and mitigation is and should be a continuously evolving process as threats and risks evolve. Bing is committed to continuously monitoring the effectiveness of the implemented risk mitigations to address issues as they arise and continue to refine mitigations, as described in the Monitoring the effectiveness of mitigations section of this report.

Conclusion

After evaluating the purpose and design of the Bing services, the key systemic risks associated with the functionality and features available on Bing, and the safety systems and other risk mitigation processes implemented by Bing, the Bing Risk Assessment team has determined that Bing has implemented mitigations that are reasonable, proportionate, and effective to address the identified risks on the service.

Key considerations in the overall conclusions include:

The nature of Bing as a search engine and the critical role that search engines play in access to information critical to functioning within society.

To uphold free expression and access to information, Bing works to avoid removing content from search results except in limited, narrow scenarios where legal demands or other important interests warrant removal. Bing has long established processes to ingest reports of concern from users and other stakeholders, including governments.

Where users are seeking low authority content that could be misleading or harmful, and Bing determines that such interventions are likely to be helpful (and not exacerbate the problem), Bing works to include additional contextual information to search results to ensure users are not harmed by content in search results. These interventions may include answers, PSAs, site-level warnings, indicators of content provenance such as NewsGuard ratings, fact checks, counternarratives for users seeking terrorist and violent extremist content to dissuade recruitment, and other high quality information to supplement what is returned in the main search results.

The low occurrence of harms on the service.

Internal estimates suggest 0.8% of queries entered on Bing are likely to lead users to unexpected problematic content. It also noted that, although Bing actions around 100 million pages of content removals every six months, the vast majority of these (~99%) are in response to copyright infringement claims and includes a small number of pages removed due to local legal obligations regarding materials that could pose risks to electoral processes, civic discourse, or public security. Given that Bing indexes hundreds of billions of webpages, this amount represents far less than 1% of the total webpages indexed. The assessment therefore determined that the probability of the identified risks occurring is low.

Reliance on Authority Principles.

Absent a clear intent to access specific content, Bing assumes an intent to find high authority results and response to search queries with the highest authority, relevant content, and grounds generative AI responses in the same principles.

The lack of user-generated content and user-to-user interaction on the service.

Since Bing generally does not allow users the ability to post content or communicate directly with other users, this limits Bing's risk exposure as to many other systemic risk areas common to online platforms, such as content virality, impersonation, or user rights regarding content removals and appeals. The risk of virality, impersonation, and user appeals is therefore low, and as such Bing has (appropriately) not dedicated significant resources to terms of use enforcement or internal complaint resolution processes.

Bing has robust processes in place that will ensure that any new features that may implicate user content will be proactively reviewed and additional risk mitigations implemented prior to launch.

The importance of allowing free and unauthenticated access to the Bing search engine and the very low number of authenticated users of the service.

As an online search engine, Bing's primary objective is to offer access to information across the web by providing relevant and high quality results in response to user queries, whether users are authenticated or not authenticated. Given the societal importance of search engines, Bing places a high priority on providing free and open access to Bing's search services. That means that a large majority of Bing users are not authenticated. This has implications on Bing's risk profile relative to the identification of minor users of the service as well as other user base demographics but helps promote user privacy, freedom of expression, and free access to information.

Bing collaborates with key stakeholders to inform risk measurement, mitigation, and broader safety considerations.

Bing works with cross-Microsoft teams to ensure compliance with internal policies and standards in a variety of high priority policy areas, such as privacy, digital safety, responsible AI, information integrity, intellectual property, diversity and inclusion, and accessibility. New product features in Bing are reviewed for compliance with each of these requirements before features are launched. Bing works with these cross-company teams of subject matter experts in order to identify new areas of concern for mitigation and to understand emerging legal requirements in the markets where it operates.

Data provided by users in Bing is handled in accordance with Microsoft privacy and security standards to ensure compliance with GDPR and other laws, and to avoid data breach.

The need to understand and address emerging threats, new risk vectors, and manifestations of harm in search.

Bing and its internal Microsoft partners work with external researchers and subject matter experts (including data sharing relationships) to supplement its internal knowledge of key subject areas, to quickly identify new threats, and to obtain feedback on the efficacy of its processes.

After evaluating the purpose and design of Bing's core search, generative AI, and ancillary search features, the key systemic risks associated with the functionality and features available on Bing, and the safety systems and other risk mitigation processes implemented by Bing, the Bing Risk Assessment team has determined that, while there are opportunities to continue to improve and mature Bing's Trust and Safety systems to address systemic risks more effectively, Bing's existing measures are sufficient and proportionate to mitigate key systemic risks on the platform.

Appendix I: Detailed Risk Assessment and Scoring Methodology

Risk Areas

Bing considered the following twelve risk areas for this Systemic Risk Assessment. In addition to risk definitions, Bing also identified common risk scenarios for each risk area to ensure adequate consideration of more specific risk or harm types within each risk area. Not all considered risk scenarios are applicable to each Bing product or service; however, they are used to explore the circumstances under which each risk outlined in Article 34 of the DSA could manifest on the service. As part of the risk assessment process, the Bing Risk Assessment team further examined whether and how each risk scenario could present on specific features of the Bing service to ensure adequate coverage of potential inherent risks with the implemented mitigations.

#	Risk Area	Risk Description
1	**Civic Discourse and Electoral Processes**	Risk that content or activities with actual or foreseeable negative effects on civic discourse and electoral processes occur on the service.
2	**Consumer Protection and Fraud**	Risk that content or activities with actual or foreseeable negative impact to a high-level of consumer protection occur on the service.
3	**Discrimination and Hate**	Risk that content or activities with actual or foreseeable negative effect on the fundamental right to non-discrimination occur on the service.
4	**Freedom of Expression and Information**	Risk that content or activities with actual or foreseeable negative effects on the fundamental rights to freedom of expression and information, including the freedom and pluralism of the media occur on the service.
5	**Human Dignity**	Risk that content or activities with actual or foreseeable negative effects on the fundamental rights to human dignity occur on the service.
6	**Illegal Content and Activities**	Risk related to the conduct of illegal activity or dissemination of illegal content through the service.
7	**Mental and Physical Wellbeing**	Risk that content or activities with actual or foreseeable serious negative consequences to a person's physical and mental well-being or in relation to gender-based violence occur on the service.
8	**Private and Family Life**	Risk that content or activities with actual or foreseeable negative effect on respect for private and family life occur on the service.
9	**Protection of Personal Data**	Risk that content or activities with actual or foreseeable negative effects on the protection of personal data occur on the service.
10	**Public Health**	Risk that content or activities with actual or foreseeable negative effects on the protection of public health occur on the service.
11	**Public Security**	Risk that content or activities with actual or foreseeable negative effects on public security occur on the service.
12	**Rights and Protection of Minors**	Risk that content or activities with actual or foreseeable negative effects on the protection of minors or respect for the rights of the child occur on the service.

Risk factors and influencers

The European Commission has identified a variety of factors that may influence the probability of risks occurring on the service as well as the potential impact of various features of the service on systemic risks. As part of its assessment, the Bing Risk Assessment team considered the impact of the following risk factors and influencers on the potential manifestation of the identified systemic risks on Bing features, the probability of the risk occurring on Bing's service, as well as the mitigations are implemented to address the systemic risks:

- The design of recommender systems and their impact on the prioritization of content that users see on the service based on the designated criteria.
- The design of generative AI features and their impact on the ability of users to accelerate the generation of potentially harmful content, as well as serve content that is potentially harmful or misleading.
- Content moderation systems - whether manual or automated - and their impact on whether, how much, and what types of content are available to users on the service. This factor can influence both the amount of harmful content that users are exposed to on the service, and it can impact user's freedom of information and pluralism.
- Terms and conditions and their enforcement and their impact on what content is available to users on the service.
- Systems for selecting and presenting advertisements and their impact on user exposure to content, user privacy and consumer protection.
- Data-related practices and their impact on consumer protection from fraud and personal data protection.
- Intentional manipulation of the service, including inauthentic use or automated exploitation, impact on consumer protection, prioritized content, and potential volume of harmful AI-generated content.
- Amplification and potentially rapid and wide dissemination of illegal or violative content via recommendation or AI-generated content and its impact on systemic risk.
- As appropriate, regional and linguistic considerations, such as the strength of content moderation in EU languages or the effectiveness of mitigation measures in EU member states.

Risk assessment inputs

This section details the evidence used to support the risk assessment scoring and rationale. The Bing Risk Assessment team collected information from a variety of sources to inform the risk assessment and has compiled this standard listing of inputs to ensure a consistent and comprehensive review of information as part of the annual assessment. The Bing Risk Assessment team gathered and reviewed the following inputs to inform and substantiate ratings assigned in the Systemic Risk Assessment.

- **Mitigation questionnaire responses:** The Bing Risk Assessment team leveraged the April 2023 DTSP questionnaire to gather a baseline of information on existing mitigations aligned with industry best practices and solicited responses to the questionnaire from internal stakeholders across the Trust and Safety, Legal, Privacy, Product, and Engineering teams.
- **Material Change Questionnaire responses**: The Bing Risk Assessment team deployed a questionnaire to each Bing product team to systematically identify changes to Bing products and services since August 2023 to be described in the 2024 Systemic Risk Assessment.
- **Policies and publications:** The Bing Risk Assessment team reviewed relevant external policies, including the content policies on the Digital Safety at Microsoft page, public practices, and other publications, including blogs, to identify additional policies and initiatives most relevant to the risk assessment.

- **Mitigation summaries:** Bing stakeholders developed brief summaries of Bing controls and mitigations specific to each risk area.
- **DSA Audit Control Inventory:** The Bing Risk Assessment team incorporated controls outlined in the DSA Audit Risk and Control Matrix and potential mitigations where relevant to systemic risks.
- **Internal consultations:** The Bing Risk Assessment team conducted in-depth workshop sessions with internal stakeholder groups to solicit more detailed information on the presentation of risks and unique mitigations relevant to various features across Bing products and services. These internal consultations included representatives from Product Teams, Bing Compliance teams, and Microsoft Defensive Intelligence Response and Transparency team.
- **External consultations:** Bing regularly engages with external stakeholders, including civil society organizations, to receive feedback on the service and to discuss best practices for addressing risk.
- **Internal policy enforcement metrics:** The Bing Risk Assessment team considered internal metrics and measurements to inform both assessment of prevalence as well as effectiveness of mitigations.
- **Transparency report metrics:** The Bing Risk Assessment team considered metrics reported through Transparency Reporting to inform the assessment of both the probability and effectiveness of mitigations.
- **External data:** The Bing Risk Assessment team reviewed a collection of social and digital media articles and conversations, including social listening reports, around the Bing service and each systemic risk area to identify trends in areas of public discourse and/or concern.
- **Authoritative sources**: The Bing Risk Assessment team reviewed publicly available sources considered reliable due to their expertise and reputation, such as regulatory sources, reputable public opinion polling, and external research, to inform the objective assessment of severity as well as the overall conduct of the Systemic Risk Assessment.

Probability

The Bing Risk Assessment team conducted a **data-driven probability assessment** to evaluate the likelihood of certain events occurring on the service absent mitigations by analyzing relevant data, including public incident data, transparency report metrics, and internal metrics.

Probability considers factors such as the inherent vulnerability and demand for the risk to occur on the service.

Assessment of **vulnerability** considers whether the perpetration of the harm on the service requires both intent and sophistication or whether it can occur without one or either. For example, can the harm accidentally occur with an average user conducting web searches online or does the harm require hacking skills with malicious intent.

Assessment of **demand** considers the volume of attempts or instances where this harm might occur. The Bing Risk Assessment team considered internal and external data samples for a data-estimated understanding of demand. For external data, the Bing Risk Assessment team reviewed public discourse on social and digital media in the EU of the considered risks in relation to Bing products and services to identify risk areas with a higher relative level of public concern. For internal data, the Bing Risk Assessment team reviewed internal metrics related to content reported by users, flagged, or removed for identified risks to identify risk areas with a higher relative of occurrence or attempted perpetration on the service.

The Bing Risk Assessment team considered the vulnerability and relative demand of the risks and compared the resulting ranking or probability to feedback from internal and external expert stakeholders, incorporating insights related to public events (for example the increased occurrence of elections this year), as well as user counts for Bing products and services, to validate the assigned probability score.

Each probability rating is assigned a score from 1 to 5, with higher scores indicating a higher likelihood of the occurrence of the event on the service absent mitigations.

Inputs that substantiate the probability assessment include:

- External data
- Transparency report metrics
- Internal data
- Internal consultations
- External consultations
- Material changes questionnaire responses

Figure 20: Probability Rating Scale

Description	Score	Rating
The risk event or circumstance is relatively certain to occur	5	Expected
The risk event or circumstance is highly likely to occur	4	Highly Likely
The risk event or circumstance is likely to occur	3	Likely
The risk event or circumstance occurring is possible but not likely	2	Not Likely
The risk event or circumstance is remotely probable	1	Remote

Severity

The Bing Risk Assessment team conducted an **objective, systems-based analysis** to determine severity, considering the complexity, scale, and gravity of impact to assign an overall severity rating.

Severity is calculated once for each risk and considers factors of complexity, scale, and gravity. Complexity is determined by the number of systems impacted by the harm. The Bing Risk Assessment team considered impact on economic, security, political, societal, environmental, and wellbeing systems. Scale is determined by whether the harm takes place at the individual, local, country, regional, or global level. Gravity considers the potential irremediability of the harm.

Figure 21: Severity Scoring Sheet

	Impact Scale				
Systems Impacted	Global	Regional	Country	Local	Individual
Economic	Low	Low	Moderate	High	High
Security	Low	Moderate	High	High	Moderate
Societal	Low	Moderate	High	High	Moderate
Political	Low	Moderate	High	High	Moderate
Environmental	Low	Low	Low	Low	Low
Wellbeing	Low	Low	Low	Low	

The Severity Scoring Sheet standardizes the consideration of the severity factors. The weighted output of the Severity Scoring Sheet corresponds with a Severity Rating on a graduated scale to avoid under-rating risks that do not have a regional or global impact but still have significant impact on several systems.

The Severity Ratings correspond to a score from 1 to 5, with higher scores indicating a higher level of Inherent Severity.

Inputs that substantiate the severity assessment include:

- External consultations
- Internal consultations
- Authoritative sources

Figure 22: Severity Rating Scale

Description	Weight	Score	Rating
Impact that could cause critical, irremediable harm, damage, or loss across several systems at a significant scale	>=243	5	Critical
Impact that could cause significant irremediable harm, damage, or loss across one or more systems at a broad scale	193 – 243	4	High
Impact that could cause some harm or disruption across one or more systems at a moderate scale, which is generally manageable or remediable	143 – 193	3	Moderate
Impact that could cause limited harm or disruption to one or more systems at a limited scale	93 – 143	2	Low
Impact that has little or no consequence on systems at any scale	<93	1	Minimal

Inherent Risk

Inherent risk is a multiplication of Inherent Probability by Inherent Severity to identify the level of risk absent mitigations.

Figure 23: Inherent Risk Rating Scale

Description	Score	Rating
A risk that is relatively certain to occur, would have a severe impact if it occurred, and requires immediate action to manage or mitigate.	>=17	Critical
A risk that is highly likely to occur, would have a significant impact if it occurred, and requires urgent action to manage or mitigate.	11 – 17	High
A risk that is likely to occur, would have a noticeable impact if it occurred, and requires some action to manage or mitigate.	6 – 11	Moderate
A risk that is not likely to occur, would have a limited impact if it occurred, and may require minimal action to manage or mitigate.	2 – 6	Low
A risk that is remotely probable, would have a minimal impact if it occurred, and may not require any action to manage or mitigate.	<2	Minimal

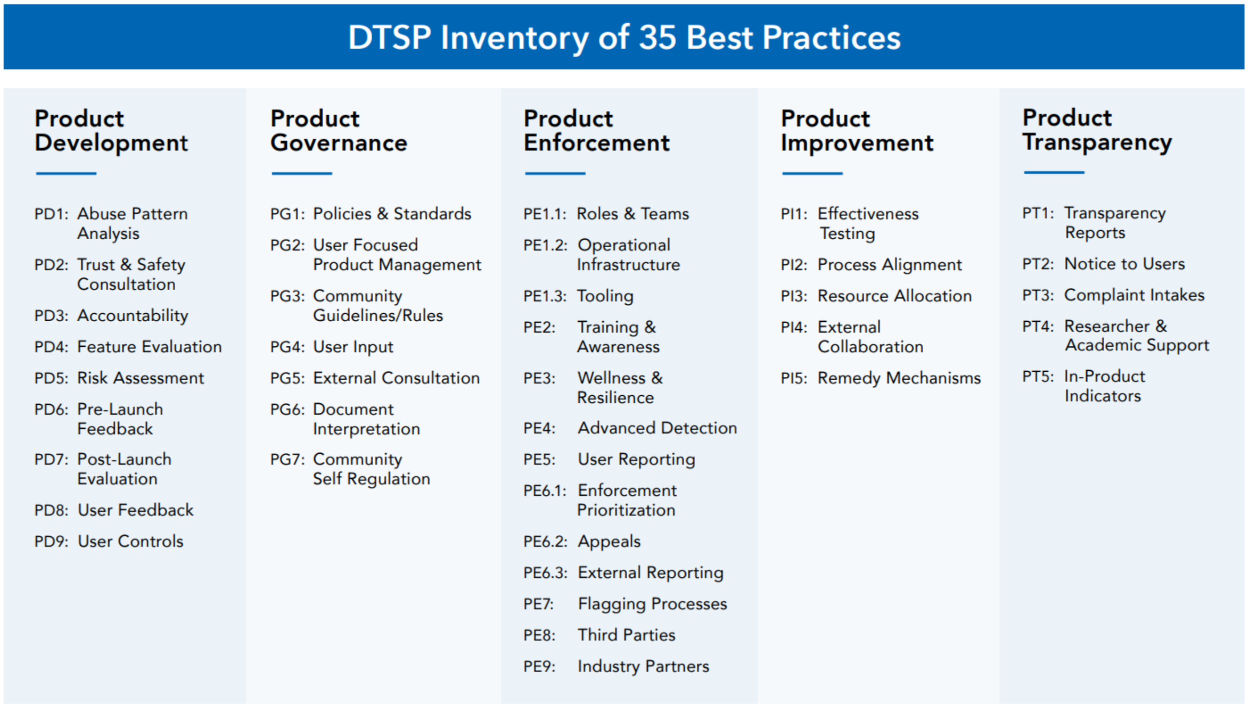
Maturity of Mitigations

The Bing Risk Assessment team then assessed the maturity of Bing's controls, safeguards, and other measures to mitigate the identified inherent risks in a reasonable, proportionate, and effective manner.

To assess the maturity of Bing's controls, the Bing Risk Assessment team collected enterprise-level mitigations, risk-specific mitigations, and feature-specific mitigations currently implemented and aligned them to the categories of the 35 Best Practices of the DTSP framework to ensure coverage across industry-standard mitigation practices. Utilizing the DTSP Framework enables Bing to set standards and enact best practices from peers across the technology industry, improved privacy protocols, balanced public discourse, and robust user protection mechanisms. Its wealth of resources empowers Bing to mitigate harmful content or actions effectively without infringing upon the freedom of expression.

On a recurring basis, Bing stakeholders complete the DTSP Safe Assessment questionnaire to identify implemented best practices within the listed 35 best practice areas. Bing does not only develop and implement mitigations aligned with this framework; rather, the framework serves as a useful assessment to identify areas in which Bing can continue to mature implemented mitigations.

Figure 24: 35 DTSP Best Practices Screenshot

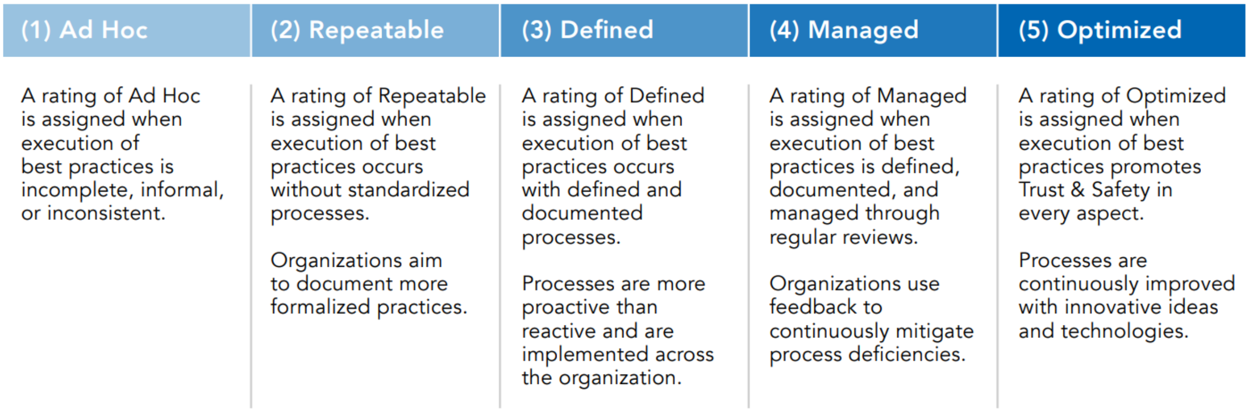


DTSP Inventory of 35 Best Practices

Product Development	Product Governance	Product Enforcement	Product Improvement	Product Transparency
PD1: Abuse Pattern Analysis	PG1: Policies & Standards	PE1.1: Roles & Teams	PI1: Effectiveness Testing	PT1: Transparency Reports
PD2: Trust & Safety Consultation	PG2: User Focused Product Management	PE1.2: Operational Infrastructure	PI2: Process Alignment	PT2: Notice to Users
PD3: Accountability	PG3: Community Guidelines/Rules	PE1.3: Tooling	PI3: Resource Allocation	PT3: Complaint Intakes
PD4: Feature Evaluation	PG4: User Input	PE2: Training & Awareness	PI4: External Collaboration	PT4: Researcher & Academic Support
PD5: Risk Assessment	PG5: External Consultation	PE3: Wellness & Resilience	PI5: Remedy Mechanisms	PT5: In-Product Indicators
PD6: Pre-Launch Feedback	PG6: Document Interpretation	PE4: Advanced Detection		
PD7: Post-Launch Evaluation	PG7: Community Self Regulation	PE5: User Reporting		
PD8: User Feedback		PE6.1: Enforcement Prioritization		
PD9: User Controls		PE6.2: Appeals		
		PE6.3: External Reporting		
		PE7: Flagging Processes		
		PE8: Third Parties		
		PE9: Industry Partners		

The Bing Risk Assessment team then mapped the mitigations to the identified risk scenarios to identify implemented mitigations to address identified inherent risks. And finally, the Bing Risk Assessment team evaluated the implemented mitigations according to the DTSP maturity rating scale. The strength of the mitigations implemented covering the identified risks and the industry best practices would correspond to a higher maturity rating; however, regardless of the strength of existing mitigations, any risk area missing mitigations across any best practice or for any identified risk scenario cannot receive a score higher than Managed maturity.

Figure 25: DTSP Maturity Rating Scale screenshot



(1) Ad Hoc	(2) Repeatable	(3) Defined	(4) Managed	(5) Optimized
A rating of Ad Hoc is assigned when execution of best practices is incomplete, informal, or inconsistent.	A rating of Repeatable is assigned when execution of best practices occurs without standardized processes. Organizations aim to document more formalized practices.	A rating of Defined is assigned when execution of best practices occurs with defined and documented processes. Processes are more proactive than reactive and are implemented across the organization.	A rating of Managed is assigned when execution of best practices is defined, documented, and managed through regular reviews. Organizations use feedback to continuously mitigate process deficiencies.	A rating of Optimized is assigned when execution of best practices promotes Trust & Safety in every aspect. Processes are continuously improved with innovative ideas and technologies.

Each Maturity Rating corresponds to a percentage score, with a lower percentage indicating a lower level of maturity.

Inputs that substantiate the severity assessment include:

- Internal consultations
- DSA control inventory
- Mitigation summaries
- Policies and publications
- Mitigation questionnaire responses
- Internal data
- Material change questionnaire responses

Figure 26: Mitigation Maturity Scale

Description	Score	Rating
A rating of Ad Hoc is assigned when execution of best practices is incomplete, informal, or inconsistent.	10%	Ad Hoc
A rating of Repeatable is assigned when execution of best practices occurs without standardized processes. Organizations aim to document more formalized practices.	30%	Repeatable
A rating of Defined is assigned when execution of best practices occurs with defined and documented processes. Processes are more proactive than reactive and are implemented across the organization.	50%	Defined
A rating of Managed is assigned when execution of best practices is defined, documented, and managed through regular reviews. Organizations use feedback to continuously mitigate process deficiencies.	65%	Managed
A rating of Optimized is assigned when execution of best practices promotes Trust & Safety in every aspect. Processes are continuously improved with innovative ideas and technologies.	75%	Optimized

Residual Risk

Inherent Risk is then multiplied against the Maturity of Mitigations score to identify a Residual Risk Rating that will help Bing to prioritize and identify which risks may require enhanced mitigation efforts.

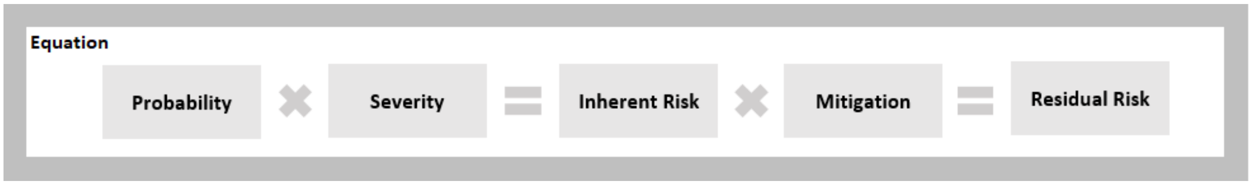
Figure 27: Residual Risk Rating Scale

A risk that is relatively certain to occur, would have a severe impact if it occurred, and requires immediate action to manage or mitigate.	>=17	Critical
A risk that is highly likely to occur, would have a significant impact if it occurred, and requires urgent action to manage or mitigate.	11 – 17	High
A risk that is likely to occur, would have a noticeable impact if it occurred, and requires some action to manage or mitigate.	6 – 11	Moderate
A risk that is not likely to occur, would have a limited impact if it occurred, and may require minimal action to manage or mitigate.	2 – 6	Low
A risk that is remotely probable, would have a minimal impact if it occurred, and may not require any action to manage or mitigate.	<2	Minimal

Sample scoring calculation

In this scoring calculation, Probability is multiplied by Severity to achieve Inherent Risk. Inherent Risk is multiplied by 1 minus Mitigation to achieve Residual Risk.

Figure 28: Systemic Risk Scoring Equation Screenshot



Equation

Probability × Severity = Inherent Risk × Mitigation = Residual Risk

Appendix II: Catalog of Mitigations by Industry Best Practices

Below is a summary of Bing's mitigations that span multiple systemic risk areas. The mitigations below have been categorized across the DTSP best practices. Although mitigations may address more than one DTSP best practice, mitigations have been categorized based on the DTSP best practice it primarily addresses.

Product Development

Abuse Pattern Analysis

Below is a summary of Bing's approach to Abuse Pattern Analysis, which is designed to develop insight and analysis capabilities to understand patterns of abuse and identify preventative mitigations that can be integrated into products.

Authority signals. Bing assigns individual websites a "QC score" to prevent low quality websites from appearing high in Bing search results. Determining the QC score of a website includes evaluating the clarity of purpose of the site, its usability, and presentation. QC scores also take into account an evaluation of the website's "authority," which includes factors such as reputation, level of distortion, and origination and transparency of the ownership. Bing collects signals from external entities like NewsGuard, GDI and fact-checked content via partnerships or Schema.org open protocol systematic intake. These signals directly identify certain low authority sites which Bing then uses to identify other low authority sites using fanout techniques.

- Bing builds and maintains automated pipelines to consume fact check labels using the Schema.org fact check open protocol.
- Bing then utilizes the fact check data to identify relevant sites. The identified sites are reviewed by human judges trained on Bing's ranking policies and principles. Once the human judges label the domains based on Bing's internal authority principles, the data along with the judgment results will feed into the domain authority dataset as authority signals.
- Bing builds and maintains automated pipelines to consume signals on authoritativeness of domains from external organizations, such as NewsGuard and GDI. These datasets will also feed into the domain authority dataset as authority signals.

Bing uses these authority signals in ranking to ensure that users are provided with high authority websites in top search results and protect users from being exposed to low authority content.

Additional Algorithmic Interventions - Defensive Search Interventions Informed by Intelligence. Although the Bing search algorithms are designed, and are continually refined, to prioritize third-party websites that that are high in relevance, quality, and credibility, Bing may in some cases identify specific threats that undermine the efficacy of these algorithms.

Bing deploys "defensive search" strategies and interventions to counteract these threats, in accordance with its search principles. Defensive search interventions may include algorithmic interventions, such as:

- QC boosts to put further weight on authoritativeness;
- Demotions of a website that has been identified as posing risks;
- Restricting the autosuggest feature in Bing to avoid directing users to problematic queries;
- In limited cases, engaging in manual interventions for specific reported issues, or in broader areas more prone to misinformation or disinformation (e.g., elections, pharmaceutical drugs, or COVID-19).

Bing actively monitors manipulation trends in identified high-risk areas and deploys mitigation methods as needed to help ensure that users are provided with high quality, high authority search results. Bing also works to identify and track nation-state information operations targeting democracies across the world and works with trusted third-party partners, including NewsGuard, the GDI, and Spanish language news agency EFE, and Agence France-Presse (AFP), to provide early indicators of narratives, hashtags, or information operations that can be leveraged to inform early detection and defensive search strategies.

Bing's defensive search interventions are designed to work across languages and markets where Bing is offered. Bing monitors language-level performance and drives continuous improvement of the effectiveness of the interventions.

Bing's defensive search efforts involve the treatment of millions of search queries and is continuously refined to address new threats, emerging narratives, and tactics used by bad actors to manipulate search rankings. For example, from July to December 2023, more than 41 million search query suggestions were suppressed as part of defensive search interventions for queries entered by users in the EU alone (including both traditional web search and Copilot in Bing). During the same period, hundreds of thousands of unique queries searched by EU users were treated with defensive search interventions.

Abuse detection in generative AI features. In addition to the above measures employed for search generally, additional abuse pattern measures are employed by Microsoft generative AI services, including red team testing, expanded reactive social listening systems, expanded and highly prominent reporting functionality, and operations and incident response.

- Microsoft uses red team testing at both the model and application layers to identify potential gaps in the safety systems.
- Reactive social listening systems: Bing maintains social listening pipelines where insights and user feedback on generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Microsoft product teams and product engineering leadership in a daily report.
- Reporting functionality and operations and incident response: Bing has set up robust user reporting and processes to review and respond to safety incidents and collect insights for further product improvement.

Trust and Safety Consultation

Below is a summary of Bing's approach to Trust and Safety Consultation, which are designed to include a Trust & Safety team or equivalent stakeholder in the product development process at an early stage, including through communication and meetings, soliciting, and incorporating feedback as appropriate.

Bing engages cross-functional teams to actively monitor and identify emerging threats through both internal and external data sources. They work together to evolve mitigation practices and adapt to new tactics aimed at circumventing Bing's safeguards, ensuring the prevention of access to potentially harmful content across various risk categories.

For changes to existing products, product teams must have conducted metrics testing of the change, including RAI metrics, and must provide data demonstrating whether the proposed change would result in improvements, regression, or no change against these metrics, and explain any additional mitigations implemented to address identified issues. The shiproom review also confirms that additional reviews such as for digital safety, security, privacy, and accessibility have been completed prior to launch. Releases are also initially limited, allowing product teams to test, measure, and implement mitigations for unintended consequences prior to general release.

Bing relies on internal and external authoritative sources of information to help improve ranking and relevance capabilities with respect to potentially harmful content/activity.

Accountability

Below is a summary of Bing's approaches to Accountability, which are designed to designate a team or manager as accountable for integrating safety feedback to the product team.

Bing engages cross-functional teams to ensure the quality of the content review processes and decisions. For example, Bing has a weekly cross-functional sync to monitor the process and quality of reviews, address any gaps, and drive improvements where needed.

In addition to the quality assurance process, Bing also completes full DPIAs for any new product or feature; the DPIA requires specific investigation of possible harms to data subjects based on processing of their personal data and documentation that Bing has identified an appropriate legal basis for processing that data under GDPR as well as to explain how Bing has mitigated possible privacy risks to users.

Generative AI features are also subject to RAI Impact assessments. To learn more, please refer to Feature Evaluation, below.

Feature Evaluation

Below is a summary of Bing's approach to Feature Evaluation, which is designed to evaluate Trust & Safety Considerations of product features balancing usability and the ability to resist abuse.

Bing relies on Microsoft's extensive cross-company compliance infrastructure to proactively review new products and features to ensure they meet Microsoft's policy commitments in key areas such as privacy, accessibility, digital safety, and responsible AI. For example, privacy reviews are an essential part of any new feature review process; implementing recommended privacy mitigations following such reviews is a requirement for launch. Microsoft has a robust privacy and security infrastructure, consisting of privacy managers who are trained in Microsoft privacy standards and relevant laws, and have access to centralized privacy specialist teams and legal support for complex or novel issues.

Similarly, generative AI features at Microsoft are subject to the RAI review process, which includes envisioning and scrutinizing the benefits and harms for stakeholders of an AI application before development or deployment. These assessments are a fundamental part of Microsoft's process to identify, evaluate, and mitigate potential risks associated with deploying AI technologies. The reviews require teams to: identify potential risks and harms, consider stakeholder impact, and propose mitigations. As part of this process teams must generally undergo “red team” testing to identify possible areas of harm, implement mitigations, and use the results of red team testing to develop scalable metrics that help measure the efficacy of these mitigations over time. The outcomes of these assessments, including identified risks and proposed mitigations, are documented and, where appropriate, shared with stakeholders to ensure transparency and accountability.

Building on the foundation of the RAI review process, Bing's meticulous approach to release readiness and ongoing monitoring of its AI features like Copilot in Bing and Image Creator from Bing heavily relies on the utilization of metrics. This process is aligned with Microsoft's RAI Standard, which requires a thorough examination of potential risks and the enactment of appropriate mitigations in the AI development and deployment stages. Metrics are central to this approach, facilitating the continuous review and update process throughout the lifecycle of AI feature creation. This ensures that any potential issues are swiftly identified and remedied.

The governance frameworks, including specific requirements for generative AI, guide the development and deployment of AI features, ensuring they adhere to high standards of responsibility. Red team testing, both adversarial and non-adversarial, further informs the development of targeted mitigations by identifying vulnerabilities.

Microsoft's ongoing commitment to responsible AI extends into post-launch, with monthly monitoring and regular metrics reviews integral to catching any issues that may have been missed during the initial release assessment. This demonstrates Bing's dedication to evolving its AI features responsibly, applying continuous improvements and leveraging metrics for rigorous testing and safety assurance.

Risk Assessment

Below is a summary of Bing’s approach to Risk Assessment, which is described as using an in-house or third-party team to conduct risk assessments to better understand potential risks. This refers to a systematic process of evaluating the potential or actual risks posed of content- or conduct-related online harms. For these purposes, Bing is focused on content- and conduct-related risks: reasonable threats to the wellbeing of people or society stemming from certain illegal, dangerous, or otherwise harmful content or behavior.

Bing conducts formal risk assessments at least once per year.

As part of the compliance processes built into the product development cycle, Bing product teams are also required to go through multiple program and compliance reviews to identify and address legal and digital safety risks, including:

- Microsoft RAI Review – Microsoft processes, programs, or tools utilizing AI, including Bing, must adhere to Microsoft's RAI Standard and undertake reviews to help ensure responsible use of AI-influenced algorithms and processes for any new product features. A central Deployment Safety Board (DSB) reviews each impact assessment, including results of red team testing and related metrics, prior to product/feature launch to ensure compliance with Microsoft’s RAI Standard.
- Privacy, Security, and Accessibility Reviews – The Privacy team evaluates products, services, and features to ensure compliance with Microsoft Privacy Standards and that personal data is protected by design and default. The Accessibility team evaluates product interfaces to ensure they are accessible to and compliant with Microsoft Accessibility Standards. The Security team evaluates products, services, and features to ensure data is secured, threats are mitigated, and compliance with Microsoft Security Standards is met.

- Legal Review – Microsoft legal teams evaluate product features and proposals to ensure compliance with applicable laws and address potential safety and legal risks posed by a product or feature. Microsoft legal teams include experts in privacy, human rights, AI, technology, marketing, IP, digital safety, consumer protection, information integrity, and litigation, among other areas and provide holistic support to Microsoft product teams.
- Microsoft Digital Safety Standard (MDSS) Reviews – Features are reviewed against MDSS to identify the potential changes to Bing's risk profile and whether additional safety mitigations are required.

Pre-Launch Feedback

Below is a summary of Bing's approach to_Pre-Launch Feedback, which is designed to provide ongoing pre-launch feedback related to Trust & Safety Considerations.

Prior to a product launch and during the software development lifecycle (SDLC), feature teams, with the support of Bing trust and safety and compliance teams, as well as cross-company subject matter expert teams, conduct a series of risk assessments as applicable to their feature, including for example, the RAI Assessment Process (incl. DSB reviews), One Compliance (1CS) assessments, Security Assessment, and Privacy reviews. In addition to risk assessments, when any feature is being created, metrics are evaluated and updated regularly. For more details, please refer to the Feature Evaluation section above.

As part of Bing's commitment to responsible AI, the process includes a formal launch readiness process to ensure that significant new generative AI features – including changes to existing products and services - are reviewed and approved by a group of key AI stakeholders across the company. It is important to note that not every feature is subject to the RAI review processes; they are specifically designed to target AI-powered features.

Additionally, every new feature in Bing and its generative AI features must go through a launch readiness review, which Microsoft refers to internally as a "shiproom" review. The shiproom process generally requires feature teams to conduct a number of different tests and evaluations and to complete corresponding documentation and review the results with leadership. This includes both offline testing and A/B testing online (often referred to as "flight" testing). For changes to existing products, product teams must have conducted metrics testing of the change, including DDR metrics, and must provide data demonstrating whether the proposed change would result in improvements, regression, or no change against these metrics, and explain any additional mitigations implemented to address identified issues. The shiproom review also confirms that additional reviews such as for security, privacy, and accessibility have been completed prior to launch. Releases are also initially limited, allowing product teams to test, measure, and implement mitigations for unintended consequences prior to general release.

Post-Launch Evaluation

Below is a summary of Bing's approach to Post-Launch Evaluation, which describes the post-launch evaluation process by the team accountable for managing risks and those responsible for managing the product or in response to specific incidents.

Ongoing evaluation of metrics. Bing teams track and monitor product quality metrics at the product/feature-level post-launch. On safety issues, Bing tracks, monitors, and reviews DDR regularly. On the areas where Bing observes regressions, Bing conducts issue spotting exercises and further invests in improvements to address the mitigation gaps.

Formal and informal feedback from stakeholders. Bing also relies on user feedback, formal or informal complaints, and feedback from internal and external stakeholders, including Microsoft policy teams, regulators, and civil society, to continue to iterate and improve on trust and safety in the Bing service.

Social listening. Bing teams maintain a social listening pipeline to gain feedback and product insights from the open Internet, from users and potential users. Such social listening reports are shared with product teams and the leadership daily to ensure the insights feed into product evaluation and further improvements.

User Feedback and Reporting

Below is a summary of Bing's approach to User Feedback and Reporting, which is designed to iterate products in light of Trust & Safety Considerations including based on user feedback and reporting or other observed effects, including ensuring that the perspectives of minority and underrepresented communities are represented.

At the bottom of each page on Bing Search, Copilot in Bing, and Image Creator from Bing, users can find a link entitled "Feedback;" users who click the link can access a free text box in which to share their views on product features or other aspects of the service. User feedback is triaged in accordance with Bing's internal guidelines and used as insights for product development.

When users have specific concerns about information they see on Bing Search, Copilot in Bing, or Image Creator from Bing, they may share their concerns through Bing's Report a Concern tool, accessible through the Feedback link available on each page of Bing. Users can address illegal content and digital safety concerns, including:

- Exposed personal or private information
- IP
- Unlawful content
- Malicious websites or spam
- Unexpected offensive or harmful material
- Issues with AI-powered features
- Any other concerns

Bing manually reviews these user reports, triages reports, sends them to the appropriate team(s) for review, and takes action where appropriate.

User Controls

Below is a summary of Bing's approach to User Controls, which refers to the practice of offering technical measures for users to control their own product experience as it relates to interactions with other users. These types of controls are especially common in social media and messaging services.

Bing provides its users options to adjust their experience in the Bing service to meet their needs; for example, users can change their SafeSearch setting (prevent the display of adult content in search results), exercise control over personal data collected by Microsoft through the Privacy Dashboard, set location/language controls, and control Bing Search inputs such as language or location.

- The SafeSearch feature allows Bing users (and for those using Family Safety features, other users in their "family" account) to control what type of adult content may appear in search results.

"Strict" mode prevents adult text and images, "moderate" (default setting in most countries) restricts explicit images, and "off" mode allows any manner of content to be displayed in search results.

- For Bing features that personalize content based on a user's interests, Bing offers controls to delete or adjust these personalized features on the applicable user interface feature page. For example, when Bing personalizes Microsoft Edge Shopping, if Bing infers from users' browsing activity that users prefer shopping at a particular store, the coupons or advertising users see might be customized to display that store's products. Similarly, when Bing personalizes users' news, if users frequently look at travel blogs and read travel articles, then users' Microsoft news feed might display more relevant news content about traveling. Users can modify their interests and control ad settings from the Microsoft Privacy Dashboard. Users can also choose whether Microsoft advertisements are personalized to their Microsoft account across devices.
- Additionally, users logged into the Bing Search service with a Microsoft Account can utilize the Microsoft Privacy Dashboard to access, export, edit and delete the personalized data they have shared with Bing and other Microsoft services. If the user is not logged into the Bing Search service with a Microsoft Account, then they can control their search history through in-product controls, such as on bing.com via Bing's settings page. Bing also offers targeted advertising and personalization controls to users.
- In the EU, any optional data collection or use via cookies or similar technologies (the primary mechanism for data tracking in an online service like Bing) requires opt-in consent and users can exercise their data subject rights via the Microsoft Privacy Dashboard leveraging the personalization & advertising toggle.
- In Bing's generative AI features, users can view individual initial conversation prompts through in-product features and can export and delete product and service usage data to delete stored conversation history through the Privacy Dashboard.

Product Governance

Policies & Standards

Below is a summary of Bing's approach to Policies & Standards, which establishes a team or function that develops, maintains, and updates the company's corpus of content, conduct, and/or acceptable use policies.

Bing MSA. Where needed, Bing provides input and feedback in support of updates to the MSA terms.

The following are Bing's key policies that mitigate systemic risks:

- **User Behavior on Bing:** The MSA governs user behavior on the service, which includes the Microsoft Code of Conduct.
- **User Behavior in generative AI Features on Bing:** Microsoft's generative AI experiences are governed by the MSA and a supplemental term (an associated code of conduct), that specifies activities that may cause a user to lose access to Bing generative AI services. Copilot AI Experiences Terms and Image Creator Terms set forth rules for use of the service to help prevent bad actors from abusing the AI-powered features to generate content that might lead to potential harms. Users that violate the terms and its code of conduct may be suspended from the service. In addition, as outlined in the "Content Moderation" section of the terms, Microsoft may

block certain prompts that violate the code of conduct or that are likely to lead to creation of material that violates the code of conduct. Additionally, generated content that violates the code of conduct may be removed. Repeated attempts to produce prohibited content or other violations of the code of conduct may result in service or account suspension. These terms are updated as needed and provided to users in relevant languages.

- **Algorithmic Prioritization of High Authority Content:** How Bing Delivers Search Results and the Bing Webmaster Guidelines contain Bing's principles for ranking and moderation of third-party content in web results and provide detailed information on the removal of content that violates laws or Bing principles.
- **Abuse/Spam:** Bing's general abuse/spam policies, detailed in Bing's Webmaster Guidelines, include details regarding prohibiting certain practices intended to manipulate or deceive the Bing search algorithms.
- **Microsoft Privacy Statement**: The Microsoft Privacy Statement outlines what personal data Microsoft collects, how it is used, purposes for which it is used, and how to access and control personal data. This includes how these components apply to specific Microsoft products like Search and Browse, cookies, Microsoft Account information, and minor's data. Additional topics about personal data are covered such as security, storage, and retention, and how to contact Microsoft. Additionally, Microsoft offers a child-friendly version of its privacy statement to better inform younger users of data practices.
- **Microsoft Advertising Policies:** Microsoft Advertising, which powers advertisements that appear on Bing, has clear and regularly enforced content policies that prevent advertising of harmful materials**.** 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 advertisements flagged to its customer support team and removes advertisements that violate its policies. Microsoft Advertising's policies prohibit political advertising. Advertisers also retain ownership and responsibility for their ad content, but the advertiser must agree to its terms when signing up for a Microsoft Advertising account. Microsoft Advertising monitors the platform and removes ads and advertisers that violate its agreement and policies. 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.

User-Focused Product Management

Below is a summary of Bing's approach to User-Focused Product Management, which institute processes for taking user considerations into account when drafting and updating relevant Product Governance.

Algorithmic Prioritization of High Authority Content: In order to provide users with the most relevant and highest authority content in search queries, Bing invests significant time and resources into ranking and relevance systems. Bing regularly reviews product quality issues, recurring trends, and emerging risks to ensure its algorithm is sufficient without subjecting the service to unnecessary censorship. In most cases, the user will be able to find the content in standard search results regardless of content being removed from generative features.

Transparency to Users: Bing provides users more detailed information on the main parameters it uses for ranking in the How Bing Ranks Your Content section of the Webmaster Guidelines, specifically calling out that in measuring the "quality" of a website, pages that call for violence, name-calling, offensive statements, or use derogatory language to make a point are generally considered low quality. Bing works to ensure these resources are available in relevant languages and markets across the EU.

No targeted advertising to users under 18 years of age: Across Bing features, authenticated users under 18 are not subject to targeted advertising. This important protection for youth will be carried forward into the new Bing features as well.

Limited data use: For users, including young people, data collection in Copilot in Bing features is used to provide the service and contextually relevant ads. Copilot in Bing features are not personalized based on past interactions with Bing features.

Data minimization: For users, long form conversational data is not associated with user identifiers and has limited retention (<30 days for any user associated information for moderation, <60 days for non-user associated content).

Child protection: Microsoft set chat outputs in Copilot in Bing to Bing's SafeSearch Strict Mode, which has the highest level of safety protection in the main Bing Search, hence helping to prevent users, including teen users, from being exposed to potentially harmful content. Parents have the ability to lock their children's accounts to SafeSearch Strict Mode from Windows Family Setting.

Copilot in Bing and Image Creator from Bing are available to authenticated users over the age of 13 (or higher in jurisdictions with a higher age of parental consent). Any user signed-in with a Microsoft account that identifies the user as under 13 years of age (or older, in jurisdictions that require parental consent for older teens) cannot access these services in an authenticated state, even in the limited version available to unauthenticated users.

- Unauthenticated users are offered limited access to the Copilot in Bing service that allows for a small number of conversations/turns per conversation.
- Image Creator from Bing is disabled for unauthenticated users.

Community Guidelines/Rules

Below is a summary of Bing's approach to Community Guidelines/Rules, which develop user-facing policy descriptions and explanations in easy-to-understand language.

Bing adheres to a set of policies and standards to ensure the products and features provided to the users are high quality, responsible, and safe. Since users do not post or share information on Bing, and instead input search queries for personal use, Bing has limited need to enforce rules as to user behavior. However, Bing has governing policies as to user content, which are enforced consistently:

- The MSA and included code of conduct govern user behavior on Bing. MSA terms are updated annually and provided to users in relevant languages. Changes are detailed in an accompanying change log.

Microsoft's generative AI experiences are governed by the MSA and a supplemental term (an associated code of conduct), that specifies activities that may cause a user to lose access to Bing generative AI services. Copilot AI Experiences Terms and Image Creator Terms set forth rules for use of the service to help prevent bad actors from abusing the AI-powered features to generate content that might lead to potential harms. Users that violate the terms and its code of conduct may be suspended from the service. In addition, as outlined in the "Content Moderation" section of the terms, Microsoft may block certain prompts that violate the code of conduct or that are likely to lead to creation of material that violates the code of conduct. Additionally, generated content that violates the code of conduct may be removed. Repeated attempts to produce prohibited content or other violations of the code of conduct may result in service or account suspension. These terms are updated as needed and provided to users in relevant languages.

- Bing leverages Microsoft-wide Microsoft Privacy Statement to ensure users understand data used in Bing and other Microsoft services. There is also a child-friendly version of the privacy statement designed to ensure younger users are able to understand Microsoft's data use and their choices. The privacy statements are made available in relevant languages, and an accompanying change log provides information about updates.
- Bing's advertising partner Microsoft Advertising has Terms of Use and Content Policies that are published and easily accessible for advertisers and Bing users. These terms are updated as needed and provided to users in relevant languages.

The primary content that appears on Bing is information from third-party websites that is linked in search results. Webmasters are not users of Bing and Bing has no contractual privity with webmasters, so Bing has limited ability to enforce rules regarding content appearing on those third-party sites. However, Bing has still taken steps to be transparent to users and webmasters as to how it ranks and moderates web content via published documents:

- Bing's Search principles: How Bing Delivers Search Results - Microsoft Support. This document describes how Bing search works, including Bing's content takedown and content moderation processes. This document is translated into relevant languages.
- Bing's Webmaster Guidelines: This Bing Webmaster Guidelines - Webmaster Tools document provides guidance to webmasters on how Bing ranks content, including when webmaster's sites may be de-ranked or not chosen for indexing due to inappropriate behavior. This document is machine-translated into relevant languages.

User Input

Below is a summary of Bing's approach to User Input, which creates mechanisms to incorporate user community input and user research into policy rules.

Bing regularly reviews user complaints via formal reporting channels as well as the user Feedback portal, which informs updates to principles and procedures. Since Bing does not allow users to post or share content on the services, Bing has limited need for "community" rules or input but does use the feedback on content appearing in search results and generative features to inform its principles and practices. See User Feedback and Reporting section.

External Consultation

Below is a summary of Bing's approach to External Consultation, which describes work with recognized third-party civil society groups and experts for inputs on policies.

Both Bing and specialized cross-Microsoft teams that support broadly across the company regularly engage with external stakeholders in areas of key policy priorities, such as terrorist and violent extremist content, information integrity and misinformation, CSEAI, responsible AI and AI-specific risks, hate speech, minors and technology, and copyright and trademark infringement, to ensure that its internal policies, practices, and standards are addressing key concerns of third-party stakeholders. These engagements can inform processes of identifying, assessing, and mitigating risks across Bing and provide greater understanding of concerns related to the Bing service and broader societal and technology related issues. Routine external engagement gives Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry.

Representative examples of external engagements undertaken during the Reporting Period include: GIFCT, GNI, the WeProtect Global Alliance, the Christchurch Call to Action, the World Economic Forum's Global Coalition on Digital Safety, the OECD, Global Project Against Hate and Extremism Project, the European Accessibility Summit, the Partnership on AI, AISIC synthetic media working group. The GDI, Truepic, Princeton University, Freedom Online Coalition, the GNI, NewsGuard, the Digital Trust & Safety Project, the Advisory Network of the Freedom Online Coalition, the United Nations' International Narcotics Control Board, ADL, AlgorithmWatch, Better Internet for Kids Forum, Family Online Safety Institute, Tech Accord to Combat Deceptive Use of AI in 2024 Elections, the COPD, EDMO Working Group for the Creation of an Independent Intermediary Body to Support Research on Digital Platforms (IIB), EU Council's "Horizontal Working Party on Enhancing Resilience and Countering Hybrid Threats" ("ERCHT"), among others.

In addition to broader dialogue, Bing may seek external feedback specifically to inform product improvements, mitigations, and policies intended to keep Bing experiences safe and reliable for users. As one example, Microsoft worked with NewsGuard to perform red team testing of Image Creator from Bing to understand the risk of misleading images generated by Image Creator from Bing. As part of their analysis, NewsGuard prompted Image Creator from Bing to create visuals that reinforced or portrayed prominent false narratives related to politics, international affairs, and elections. Based on these evaluations, Microsoft improved existing Image Creator from Bing mitigations.

Bing also engages with key external stakeholders and industry members through the COPD (including the Elections Working Group and generative AI & Disinformation subgroup, which Microsoft sub-chairs).

Moreover, Bing regularly meets with regulators around the world, including the European Commission and EU member state Digital Service Coordinators, to understand key concerns, share information, and incorporate feedback into product design and safety systems as appropriate.

For example, Bing and Microsoft representatives presented information on mitigations related to elections, both before and after EU elections, to the European Commission, digital service coordinators, and other member state authorities and has participated in EU elections tabletop workshops organized by the Commission dedicated to assessing risk scenarios concerning elections in the European Union. As another example, Bing responded to a consumer protection report released by a (non-EU) government agency concerning the ability of bad actors to use search engines to find websites offering stolen credit card details and other sensitive information that criminals can use to defraud people. As part of its response, Bing undertook measures to address issues raised in the reported, including removal of certain identified websites for policy violations and adjustments to related search suggestions and committed to continue undertaking efforts to address risks that search results can lead users to sites involved in illicit activities, such as credit card fraud. Bing has engaged with other government agencies on issues pertaining to child safety, consumer protection, CSEAI, and NCII and these engagements are important tools to inform the design and performance of Bing safety systems. Bing has also responded to regulatory feedback on product recommendations in Copilot in Bing and made adjustments following discussions.

In addition to direct engagements, Bing provides various tools and data sources to third parties to support transparency, research, and examination of systemic risks. Bing recently formalized its Qualified Researcher Program to enable EU researchers to easily request access for publicly accessible Bing data from a singular landing page. In addition, Bing is continuing to evaluate additional data sources and tooling to support the research community and looks forward to supporting the DSA vetted research program as it is formally established.

Document Interpretation

Below is a summary of Bing's approach to Document Interpretation, which documents for internal use the interpretation of policy rules and their application based on precedent or other forms of investigation.

Bing's defensive strategies and principles are continually updated in response to learnings from prior escalations and investigations. Bing works with both internal and external stakeholders such as Legal, Digital Safety Office, Office of Responsible AI, Public Policy, Communications, other product orgs, external researchers, and civil organizations, on high-risk cases and edge cases where there is complexity in the application of principles or gaps. Such escalations are discussed and analyzed in a weekly deep-dive among product and legal stakeholders and, where applicable, policy updates may be made by Bing to ensure that its principles address the prevalent issues in the online safety space. Bing may update the internal strategies, guidelines, and external user-facing documents to capture the updates.

Community Self-Regulation

Below is a summary of Bing's approach to Community Self-Regulation, which facilitate self-regulation by the user or community to occur where appropriate, for example by providing forums for community-led governance or tools for community moderation and find opportunities to educate users on policies, for example, when they violate the rules.

Considering the nature of Bing as a search engine, self-regulation is not a frequently used mechanism. However, individuals are able to manage their own experience through user controls. See User Feedback and Reporting section.

Product Enforcement

Roles and Teams

Below is a summary of Bings approach to Roles and Teams. This constitutes roles and/or teams within the company accountable for policy creation, evaluation, implementation, and operations.

- Bing's Responsible AI Champ is accountable for implementing the responsible AI practices across Bing products and features. The process includes assessing the risks and possible harms associated with a product, capturing system details and safety concerns, and proposing appropriate mitigations.
- Bing's Digital Safety Program Owner oversees the digital safety practices across Bing products and

features, including conducting product design safety reviews, making safety mitigation recommendations, facilitating moderation ops process building, and consulting on measurement and monitoring.
- Bing's Compliance team is responsible for privacy, security and accessibility reviews and building related policies. This team works with respective centralized Microsoft teams (privacy, security, and accessibility) responsible for the creation of company-wide policies and operations.
- Bing's Defensive Intelligence Response and Transparency (DIRT) team is accountable for search principles creation, evaluation, and operations.
- Bing's Customer Services and Support team is accountable for reviewing and triaging user reports.
- Bing's Content Moderation Ops team is accountable for reactive defensive reviews and labeling.
- Bing's Defensive Search team is accountable for implementing algorithmic interventions and metrics monitoring.
- Bing's legal team provides support to ensure search principles are aligned with legal requirements and central Microsoft policies and standards.
- Microsoft's cross-company subject matter expert teams, policy teams, and government affairs teams are accountable for development and ongoing support of cross-company policies in specific high profile policy areas, which inform Bing's policies.

Operational Infrastructure

Below is a summary of Bing's approach to Operational Infrastructure. Bing developed and reviewed operational infrastructure facilitating the sorting of reports of violations and escalation paths for more complex issues.

Bing has a defined set of review protocols, including policies for handling reported violations and escalations in Bing's Customer Protection Online Safety Playbook. Bing's Content Moderation team adheres to the guidelines in the playbook in handling violations and escalations.

Bing's Customer Service and Support (CSS) team is the main team that handles the Trust & Safety tickets as the entry point, reviews and takes actions or triages to relevant teams where appropriate.

- Scope: legal removal orders, government takedown requests, copyright infringement removal requests, Right to be Forgotten requests, and user reports concerning the content appearing in Bing Search, Copilot in Bing and Image Creator from Bing.
- The CSS team is accountable for completing the initial review within 3 days. Detailed Key Performance Indicators (KPI) as follows:
  - Minutes Per Incident (MPI): 4.0
  - Service Level Agreement (SLA): 98% within 24 hours
  - Time to Acknowledgement (TTA): <24 hours
  - Time to Resolution (TTR): <= 3 days
- For the tickets that are within the review scope of CSS, the human reviewers review the tickets that are triaged in an internal ticket management tool ServiceNow (SNOW), and where appropriate, take actions in the Content Moderation Portal.
- For the tickets that are outside the review scope of CSS, the human reviewers triage the tickets to other teams including Bing product teams, legal team, privacy team, etc.
- The human reviewers receive an onboarding training of principles and internal policies, guidelines and processes as well as tooling at the time of ramping up, and additional training when there are policy updates.

Bing's Content Moderation Ops team is accountable for handling the defensive intervention related jobs that get triaged by the CSS team as described above.

- Bing's Content Moderation Ops team reviews jobs that require judgment based on defensive principles and guidelines and handles copyright claims.
- The Content Moderation Ops team's vendor team lead has weekly syncs with a team within Bing, which is accountable for policy creation and evaluation, to ensure consistency in the policy understanding and enforcement.
- The human reviewers receive an onboarding training of principles, guidelines and processes as well as tooling at the time of ramping up, and additional training when there are policy updates.

Microsoft Advertising is accountable for ensuring the advertising content served is legal, clear, truthful, and accurate.

- Advertising employs a robust filtration system to detect robotic traffic and other harmful cyber-attacks.
- Microsoft Advertising has several teams of security engineers, support agents, and traffic quality professionals dedicated to continually developing and improving the 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.

Tooling

Below is a summary of Bing's approach to Tooling. This determines how technology tools related to Trust & Safety will be provisioned (i.e., build, buy, adapt, collaborate).

Bing uses various tools to handle flagged and removed content, and investigations. Bing has shipped the Unified Defensive Document Model through the OutboardDU platform, enabling them to stamp problematic defensive documents across defensive threats, adult, and racy scenarios for Bing Web Index (400B) universally.

Additionally, Bing uses internal Microsoft tools and technologies for trust and safety including PhotoDNA which is a hash-matching technology. Bing employs this technology to identify duplicates of verified CSEAI so it can be removed from the Bing index, and also to block new copies of these images from being uploaded.

Bing uses Report a Concern to allow users to submit reports to Bing. This webform is a centralized intake portal for Bing users to submit trust and safety related concerns to Bing. The tool intakes the information submitted by users, labels accordingly, and sends the tickets to relevant internal tools for reviewing and processing. For more information on Report a Concern, see the User Reporting section.

Bing utilizes its own Content Moderation Platform (CMP), which is an internal tool for implementing content moderation decisions, such as URL exclusion and query treatment. Bing's Content Moderation team uses CMP to effectively take actions on URLs/queries based on the Bing search policy and implement the decision in the backend. For more information, see the Civic Discourse and Electoral Processes section of the Risk Assessment Summary.

For responding to tickets, Bing uses SNOW. SNOW is a ticket management tool. Bing uses SNOW to effectively manage trust and safety tickets and track and monitor SLAs.

Lastly, Universal Human Relevance System (UHRS), is a crowdsourcing platform that supports data labeling. Bing uses UHRS for making judgments on the authority of domains.

Training & Awareness

Below is a summary of Bing's approach to Training and Awareness. Formalized training and awareness programs to keep pace with dynamic online content and related issues to inform the design of associated solutions.

Bing's Content Moderation Ops team carries out training to ensure alignment on policy, labelling, and enforcement.

- The team lead of the Content Moderation Ops team has weekly syncs with cross-functional teams within Bing to discuss work items, edge cases and address policy questions, to ensure alignment and consistency between the policy team and operations team and discuss emerging trends observed.
- Content reviewers receive extensive training on Bing principles and internal policies, including the rationale behind them and how to apply them accurately. Decisions are periodically checked to ensure the policies are being applied consistently. Ongoing coaching and training are provided as sometimes laws change, new types of harms surface, or policies need to adapt.
- For high-consequence harms, like child sexual exploitation and abuse, specialized teams receive additional focused training and wellness resources.

Bing carries out training to ensure that human reviewers are able to meet Bing's standards regarding labelling quality, understanding of policy, and accuracy of labelling.

- Bing's Product Managers have weekly deep-dive sessions with cross functional teams within Bing to discuss labelling pipelines, methodology, etc. to ensure alignment and consistency and to identify new risk areas, gaps.
- Bing conducts onboarding training on the labelling guidelines as well as ongoing auditing to ensure auditors receive sufficient coaching on the policy interpretations.

Microsoft requires employees to undergo regular training in key areas of legal and policy compliance, including privacy, security, and standards of business conduct. Relevant product managers or "champs" receive additional training and regular communications from expert teams in their areas of focus, such as for digital safety, privacy, or responsible AI.

Wellness & Resilience

Below is a summary of Bing's approach to Wellness & Resilience. Bing invests in wellness and resilience of teams dealing with sensitive materials, such as tools and processes to reduce exposure, employee training, rotations on/off content review, and benefits like counseling.

Bing leverages the investments made by Microsoft in content reviewer training and wellness - from top notch medical benefits to a robust Employee Assistance Program (EAP) service, as well as comprehensive fitness and mental health programs on site and virtually. Content Moderation and review teams have guidelines of what not to review, how to address the content they accidentally view, and the ability to opt-out of work related to mental health and wellness. Furthermore, general Microsoft health and wellbeing policies apply to ensure employees' holistic wellbeing – mental, emotional, physical, and financial - is supported.

Algorithmic Mitigation

Below is a summary of Bing's approach to Algorithmic Mitigation, where feasible and appropriate, to identify areas where advance detection, and potentially intervention, is warranted.

Bing must rely on a combination of proactive and reactive processes and procedures to help minimize the likelihood that users (or indirectly, social institutions) are not harmed or misled by materials that are returned in search results. In general, and with limited exceptions, Bing does not proactively scan for, monitor, or identify illegal or otherwise problematic materials in search results, as the context in which the material appears is relevant to the determination of whether or not it is illegal. However, in limited cases, Bing may use advance detection and intervention to address significant legal or policy violations, such as CSEAI. In the context of Bing's generative AI features, Bing may also use advance detection to triage potential code of conduct violations.

1. Proactive Removal of CSEAI

One exception where Bing proactively removes illegal content from entering the search index is with regard to CSEAI. The production, distribution, and access to CSEAI materials is universally condemned as a major societal harm and is generally illegal in most jurisdictions. Bing takes a more proactive approach to tackle this issue by preventing pages from entering the index that have been reviewed by credible agencies or identified using Microsoft PhotoDNA hash-matching technology and found to contain or relate to the sexual exploitation or abuse of minors, and regularly reviews content in the index for newly identified CSEAI.

Bing removes pages from its index that have been identified by the IWF (UK), NCMEC (US), and FSM (Germany) as, in their good faith judgment, hosting or providing access to child sexual abuse material. Removing these links from displayed search results does not block the materials from being accessed on the web or discovered through means other than Bing, but it does reduce the ability of those who would seek it out or profit from it by removing it from the Bing Search index.

The Bing Visual Search feature allows users to use an image as a query to search for similar images. Bing also uses the hash-matching technologies PhotoDNA and MD5 to detect matches of previously identified CSEAI in these images provided by users. In the context of the immediate search, the use of these technologies furthers Bing's goal to avoid inadvertently surfacing potentially harmful web content to users. More broadly, images uploaded to Bing Visual Search typically contribute to training Bing's image-matching algorithms; by scanning images Bing helps to ensure that CSEAI is not included in training data.

2. **Additional Safety Interventions for Copilot in Bing and Image Creator from Bing**

Copilot in Bing is designed to provide responses supported by the information in web search results when users are seeking information. Image Creator from Bing is designed to help users generate digital images from text-based prompts. Microsoft has implemented additional filtering and classifiers to prevent chat responses from returning what Bing considers "low authority" content as part of an answer and to help address impermissible content and behaviors.

Copilot in Bing and Image Creator from Bing deploy mitigations in the form of "classifiers" and "metaprompting" to help reduce the risk of certain harms and misuse of the services since LLMs can potentially generate problematic content. ***Classifiers*** classify text to flag different types of potentially harmful content in search queries, chat prompts, or generated responses. Bing uses AI-based classifiers and content filters, which apply to search results and relevant features; Bing designed additional prompt classifiers and content filters specifically to address possible harms raised by these generative AI features. Flags lead to potential mitigations, such as not returning generated content to the user, diverting the user to a different topic, or redirecting the user to traditional Search. For example, Copilot in Bing's classifiers can detect jailbreak prompts, which are prompts with sole purpose of circumventing Copilot in Bing's safety systems to create content in violation of Copilot terms and restricts model responses. Another example is, in Image Creator from Bing, Microsoft has put controls in place that aim to limit the generation of harmful or unsafe images. When Image Creator from Bing's system detects that a potentially harmful image could be generated by a prompt, it blocks the prompt and warns the user.

In order to block generation of content based on user inputs that include certain names or terms, both Copilot in Bing and Image Creator from Bing use blocklists to block such content. For instance, Image Creator from Bing uses blocklists that contain the names of public figures, e.g., celebrities and politicians. Microsoft also allows living artists, celebrities, and organizations to request that their names and/or brands be added to these blocklists, meaning that user prompts using their names or brands will not generate an image based on those blocked names or brands. Bing also has additional blocklists that contain terms that are likely to result in generation of content that is strictly protected by law (e.g., child sexual exploitation and assault imagery), politically charged, or degrading or otherwise sensitive. In addition to these "hard" blocklists, Image Creator from Bing has a "soft" blocklist of terms that are lawful but have the potential to generate harmful content. For these terms, Microsoft has established a process to label the generated images and filter out the harmful content before it is shown to the user.

While Bing deploys the above-mentioned mechanisms to help reduce harms, Bing is also continually working to ensure that its generative features do not overblock outputs so that users are able to access the information they seek. Bing measures and monitors conversation metrics to improve the interventions to balance the harm prevention and provide users with useful information.

Microsoft is harnessing the data science and technical capabilities of its AI for Good Lab and Microsoft Threat Analysis Center teams to better detect abusive synthetic media and other harmful content on the Internet. Furthermore, Microsoft's Digital Crimes Unit is investing in new threat intelligence work to pursue the early detection of AI-powered criminal activity.

User Reporting

Below is a summary of Bing's approach to User Reporting, to implement methods by which content, conduct, or a user account can be easily reported as potentially violating policy (such as in-product reporting flow, easily findable forms, or designated email address).

Microsoft has designed Bing to provide multiple avenues through which users of Bing's search services can report concerns. Refer to User Feedback and Reporting section.

Bing also enables users with a report functionality across its features. For instance, in Maps, users can report images and locations they deem to be harmful or false. These reports go through a review process as do any user suggested locations in Map Builder, edits to businesses, and posted photos to businesses. Users are not able to write reviews.

Enforcement Prioritization

Below is a summary of Bing's approach to Enforcement Prioritization, which seeks to operationalize enforcement actions at scale where standards are set for timely response and prioritization based on factors including the context of the product, the nature, urgency, and scope of potential harm, likely efficacy of intervention, and source of report.

Content identified by reporters and classifiers as possibly in violation of laws or Bing policy is categorized and prioritized based on the reporting reason selected by the reporter or associated classifier, which also dictates the time period in which the content should undergo review. Content is prioritized by legal SLAs are the highest priority, second priority is key priority policy areas (e.g., CSEAI removal), and third priority is significant external visibility of issue. Additionally, other considerations such as the content's language, the region from where the content originated, and the media type also affect the prioritization of the review process. Most content issues are resolved within 24 hours regardless of prioritization.

Any ads and associated components within an advertisers account which violates policies within the following categories (Ad Requirements, Disallowed Content, Extensions, IP, Legal and Privacy, Media Formats, Product Ads, Relevance and Quality, Restricted Content), will be subject to a "three strikes" enforcement penalty. If an advertiser is under a strike penalty, they will not be able to log into their account through Editor or the Mobile app but can access through the Web UI. Microsoft Advertising determines strikes by policy categories for which relate to the violation, or any violation that may poses a risk to the safety or security or Bing customers, or users. Strike one will result in Microsoft Advertising suspending ads and associated components of the advertiser. Strike two, Microsoft Advertising will prevent the Manager Account from creating new accounts for the duration of the penalty. At strike three, advertiser's accounts will be suspended.

Appeals

Below is a summary of Bing's approach to ensuring appeals or other appropriate access to remedies are available.

Given that Bing users do not host, post, or share content on the service, there is no need for Bing to establish appeal processes for user content removals.

In the generative AI experiences, Bing may take action to restrict access to the generative AI services where a user has engaged in serious violations of the code of conduct, such as repeated attempts to bypass safety protections or attempts to generate CSEAI or other extremely harmful content.

- Users of Copilot in Bing will receive an in-product notice informing them of the restriction, upon which they may appeal that decision, as described in the Copilot Terms.
- In Image Creator from Bing, if a user's prompts are repeatedly blocked due to policy violations, the user is automatically suspended from image generation for a period of time or permanently. Users can appeal the permanent ban, as described in the Image Creator Terms.

Such appeals are reviewed by human judges based on the terms and decisions would be made either to uphold the original restrictions or overturn the decision and re-instate user access to the services.

While not related to user content, Bing also offers webmasters who have created Bing Webmaster accounts to appeal decisions regarding content downranking or removal. Webmasters may appeal such decisions via Bing Webmaster Tools.

Advertisers may appeal disapproved ads or components, any strike, or an immediate suspension by contacting Microsoft Advertising Support within six months of the disapproval or suspension decision. If they redress the violation, Microsoft will lift applicable penalties from your account.

External Reporting

Below is a summary of Bing's approach to External Reporting. Appropriate reporting is done outside the company, such as to law enforcement, in cases of credible imminent threat to life.

If Bing receives requests to remove content from individuals, businesses, and governments, in limited cases, where quality, safety, user demand, relevant laws, and/or public policy concerns exist, Bing might remove results, inform users of certain risks through PSAs or warnings, or provide users with options for tailoring their content. Bing limits removal of search results to a narrow set of circumstances and conditions to avoid restricting Bing users' access to relevant information.

Flagging Processes

Below is a summary of Bing's approach to Flagging Processes. Bing ensures that relevant processes exist that enable users to "flag" or report content, conduct, or a user account as potentially violating policy, and enforcement options on that basis.

Bing users are not able to post or share content in the service, so there are few scenarios where user conduct or content would raise concerns that would require reporting. As described above, Bing does provide users with the ability to report concerns with content appearing in search results or generative AI features and provides a "feedback" link on every page to enable simple reporting of user concerns.

Third Parties

Below is a summary of Bing's approach to working with Third Parties. Bing works with recognized third parties such as qualified fact checkers or human rights groups to identify meaningful enforcement responses.

Because Bing does not currently allow for publication or promotion of user-generated content, it does not have much opportunity to engage in enforcement actions against users in the same manner that a social network may utilize such resources. Bing takes action when users attempt to bypass safety protections in generative AI or visual search to try and create or find CSEAI.

While not related to enforcement actions against user content, Bing does work with globally recognized independent organizations to further refine search results in key policy areas such as mis/disinformation, terrorist and violent extremist content, fraud, data security, and child safety. Bing carefully vets its partnerships to ensure that such relationships do not inadvertently introduce bias into Bing systems and ensures any third-party vendors or partners who are given access to Bing data are upheld to standard data processing practices via contract e.g., third parties acting as Processors for Bing are required to contractually confirm compliance with Article 28 of the GDPR and join the Microsoft Supplier Security and Privacy Assurance (SSPA) program, which holds vendors to a high level of accountability.

For instance, Travel works with third parties for Travel Stories, and they have separate content moderation and quality assurance processes that are upheld to Bing's standards. For content moderation, licensors represent, warrants and covenants that the data content is not misleading, false, libelous, defamatory, obscene, unlawful, or injurious to any third-party. Bing undertakes additional quality checks of Travel

Stories content, and Bing can report issues to the vendor to remediate. Some of these data content issues include inaccurate URLs, dead links, adult, risqué or deleterious data content, copyrighted data content, data content that violates Microsoft policy, or illegible characters. For first priority items, third-party response time is less than 24 hours and resolution time is less than 3 days and corrected feed or Application Programming Interfaces (API) content is delivered within 5 days after first reported.

Bing manually monitors intelligence via threat intake channels, such as information received from trusted third-party partners including the IWF, the Institute for Strategic Dialogue, NewsGuard, EFE, the GDI, and other external and internal data sources, to identify active threats. The results of this monitoring are logged, along with the associated domains/URLs and keywords, in a central repository. These insights are also escalated to the product teams to feed into the development of classifiers to further mitigate identified risks.

Microsoft also partnered with NewsGuard to perform red team testing of Image Creator from Bing, to understand the risk of misleading images generated by Image Creator from Bing. As part of their analysis, NewsGuard prompted Image Creator from Bing to create visuals that reinforced or portrayed prominent false narratives related to politics, international affairs, and elections. Based on these evaluations Microsoft improved the existing Image Creator from Bing mitigations.

Industry Partners

Below is a summary of Bing's approach to working with Industry Partners. Bing works with industry partners and others to share useful information about risks, consistent with legal obligations and security best practices.

Bing collaborates with industry partners to better identify risks and share best practices. For example, Microsoft is part of the GIFCT, the GNI, Tech Accord, and the DTSP.

Other industry groups include, but are not limited to:

- Participation and leadership in industry associations and organizations:
    - Microsoft is a founding and active member of the C2PA and is currently a co-chair.
    - Microsoft and key members of the Bing Search team are also involved in the Partnership on AI (PAI) to identify possible countermeasures against deepfakes and has participated in the drafting and refinement of PAI's proposed Synthetic Media Code of Conduct. The proposed Code of Conduct provides guidelines for the ethical and responsible development, creation, and sharing of synthetic media (such as AI-generated artwork).
    - Microsoft is also a founding and active member of the Tech Coalition, focused on facilitating industry cooperation to address online child sexual exploitation and abuse, and the Global Internet Forum to Counter terrorism, which focuses on preventing terrorists and violent extremists from exploiting digital platforms.

- Regular participation in industry legislative working groups in connection with the COPD, working groups related to Australia's e-safety law and COPD, and others.

Microsoft has extended its support to the Verifee Project in response to disinformation campaigns targeting Slovakia. Launched in the fall of 2023, Verifee is a browser extension and digital literacy program that uses AI to transparently rate the credibility of online news articles. With the capacity to process large amounts of data, Verifee is able to recognize manipulative elements in the text.

Additionally, together with Open AI, Microsoft has established a Societal Resilience Fund to promote AI education and literacy, supporting organizations like AARP and International IDEA.

Product Improvement

Effectiveness Testing

Below is a summary of Bing's approach to Effectiveness Testing, which is designed to develop assessment methods to evaluate policies and operations for accuracy, changing user practices, emerging harms, effectiveness, and process improvement. As a starting point, one would evaluate the controls currently in place to determine whether they are operating as intended to mitigate the intended risk. Control strength evaluation includes testing both the design and effectiveness in execution.

Metrics Monitoring – DDR: Bing uses DDR as a primary means to measure the efficacy of implemented safety mitigations. DDR measures the likelihood of content violating Bing principles and internal policies appearing in search results or generative AI outputs across sets of sample queries.

- The metrics are reviewed monthly in shiproom meetings with Bing leadership~~s~~ and feature teams; when regressions are observed, teams conduct issue identification and implement fixes or drive improvements.
- DDR measurement is also a part of pre-launch assessment.
- Bing conduct analysis on DDR for each DSA systemic risk as an underlying analysis for the formal risk assessment.

Metrics Monitoring – Grounding: Microsoft works to systematically measure the efficacy of Copilot in Bing grounding responses to factual information in web search results.

Red Team Testing: Microsoft uses red team testing at both the model and application layers. The Microsoft RAI Standard requires the evaluation (including red teaming and systematic measurement) of AI systems to map and measure a variety of risks, including the risk that the system will generate harmful content. Whether testing at the model or application layer, the AI red teaming processes generally involve taking some or each of the following steps before, during, and after each round of testing:

- Develop a testing plan, which outlines the risks and features the red team will assess, and the attack approaches to take, as part of their work. Different features will have different red team requirements in light of the novelty and risk of the feature.
- Identify the ideal composition of red teamers in experience, demographics, and expertise across disciplines (for example, experts in RAI, privacy, or security) for the particular domain. Microsoft also looks for red teamers with both benign and adversarial mindsets—i.e., both those with security-testing experience, and also ordinary users of the application who have not been involved in the development of the feature being tested.
- Assign red teamers to particular harms or product features—for example, security subject matter experts may be assigned to probe for jailbreaks. Teams may also assign red teamers to specific features in an application to ensure coverage of the entire application.
- Consider how much time and effort each red teamer should dedicate. For example, when testing

for benign scenarios, less time may be needed than when testing for adversarial scenarios.

- Provide instructions to red teamers on how to test, to help them understand the purpose of the red teaming, the product that will be tested, the kinds of issues to test for, and how to record results.
- Iterate. When testing for known harms or the effectiveness of mitigations, it is not unusual to identify new harms. Teams integrate these into testing.
- In some cases, Microsoft might invite external parties to also red team-test (Such as the red team testing performed by NewsGuard for Image Creator from Bing, which is described in the Third Parties section of the Appendix II).

User feedback and reports: Bing regularly reviews user feedback and reports to not only triage concerns but to take insights for continuous product improvement and evaluate the effectiveness of controls in place to protect users.

Social listening: Bing maintains social listening pipelines where insights and user feedback on Bing's generative AI features are collected from the open Internet. These insights and user feedback are manually reviewed by humans, analyzed daily, and shared across Bing's product team and product engineering leadership in the daily report. These reports inform Bing as to the public perception of their mitigations and serve as barometer on topics that are concerning Bing's operations.

Process Alignment

Below is a summary of Bing's approach to Process Alignment, which is looks at establishing mechanisms for ensuring that policies and operations align with the 5 DTSP Commitments.

The MDSS reflects a company-wide commitment to addressing content and conduct-related risks in each of the 5 DTSP commitment domains. Product leads and legal representatives utilize the MDSS to identify and evaluate content risks and determine appropriate content moderation frameworks to manage those risks. The provisions of the MDSS can also help Bing executives and other digital safety stakeholders to identify operational and resourcing needs the product, or feature will need to maintain safety thresholds and implement necessary improvements. Both product leads and digital safety stakeholders are responsible for documenting their policies, terms of use, community guidelines, enforcement mechanisms and relevant transparency documentation to comply with MDSS.

Additionally, Microsoft's Digital Safety Leadership Council brings together Digital Safety Accountable Leaders from across the company to share best practices for implementing safety by design. Digital Safety Program Owners are responsible to execute MDSS goals and requirements for their individual products and services and socialize best practices with their work groups.

Resource Allocation

Below is a summary of Bing's approach to Resource Allocation, which is designed to use risk assessments to determine allocation of resources for emerging content- and conduct-related risks.

As mentioned in the Product Development section above, Bing regularly engages in risk assessment and evaluates the need for resource allocation to address emerging concerns. Bing sets formal OKRs every six months based on review of metrics and areas for improvement and considers engagement across the organization as well as investment levels as part of the OKRs. On a monthly basis, Bing senior leadership aligns with cross-functional teams to ensure resources are allocated appropriately, especially for key areas of content concern. Finally, Bing teams send daily updates to product leadership on key areas of content and conduct concern to ensure appropriate prioritization particularly for improvements to generative AI features.

External Collaboration

Below is a summary of Bing's approach External Collaborations, which are designed to foster communication pathways with users and other stakeholders (such as society and human rights groups) to update on developments and gather feedback about the social impact of the product and areas to improve.

See External Consultation section.

Remedy Mechanisms

Below is a summary of Bing's approach to Remedy Mechanisms, which are designed to establish appropriate remedy mechanisms for users that have been directly affected by moderation decisions such as content removal, account suspension, or termination

Due to the nature of Bing as a Search engine, it generally does not host user-generated content and therefore does not remove user accounts and content in the same manner that a social media platform would. However, Bing has clear policies and reporting processes for users where applicable. Refer to Appeals section above.

Product Transparency

Transparency Reports

Below is a summary of Bing's approach Transparency Reports, which are designed to publish periodic reports including data on salient risks and relevant enforcement practices, which may cover areas including abuses reported, processed, and acted on, and data requests processed and fulfilled.

Bing is transparent about its policies and actions with respect to user and webmaster content and makes these documents available in EU member state languages. Microsoft's Reports Hub is a publicly available source where users can access various transparency reports in areas where Bing reports on various practices. These key areas include RAI, content removal requests (such as copyright or digital safety), privacy (such as "Right to be Forgotten") and security (such as Government Requests.) The Reports Hub contains additional transparency reports such as jurisdictional, community and privacy and security transparency reports that users can easily access. As a result of the conduct of the Systemic Risk Assessment, each year Bing identifies specific areas for focused enhancements in the coming year.

The Microsoft Bing Ad Library is another initiative aimed at enhancing user understanding and control over the advertising experience. This resource allows users to delve into the specifics of the ads displayed on Bing, including details about the advertisers, their expenditure, and the rationale behind the presentation of certain ads. The "Why This Ad" feature further demystifies the ad selection process by explaining the relevance of ads based on individual online behavior and interests, thereby offering users greater insight and authority over their digital environment.

Additionally, Bing is a signatory to the COPD, which requires biannual reporting and compliance with anti-disinformation measures to improve Bing's ability to fight misinformation across the EU, including commitments to providing users with indicators of content provenance, fact checks, and researcher access to data.

Notice to Users

Below is a summary of Bing's approach Notice to Users, which is designed to provide notice to users whose content or conduct is at issue in an enforcement action (with relevant exceptions, such as legal prohibition or prevention of further harm).

In many cases, Bing's moderation decisions will not affect user content since Bing does not allow users to upload or share content on the service, and there are limited cases in which a user's activities on the service (such as entering a prompt in Copilot in Bing) could trigger enforcement actions.

- Webmasters (who are not Bing end users) have the ability to see how their content has been indexed by Bing in their Bing Webmaster Dashboard.
- In the case of Terms of Use violations in generative AI features, users receive in-product notice informing them of the access restrictions or suspensions.

Complaint Intakes

Below is a summary of Bing's approach to Complaint Intakes, which are designed to log incoming complaints, decisions, and enforcement practices in accordance with relevant data policies.

Bing users can report concerns about content appearing on the service directly through any feature across the Bing service. See User Feedback and Reporting section.

Researcher & Academic Support

Below is a summary of Bing's approach to Researcher & Academic Support which are designed to create processes for supporting academic and other researchers working on relevant subject matter (to the extent permitted by law and consistent with relevant security and privacy standards, as well as business considerations, such as trade secrets).

Microsoft's commitment to fostering a trustworthy digital ecosystem is evident through its strategic partnerships, innovative research initiatives, and dedication to public well-being. Microsoft collaborates with esteemed research and academic institutions to help ensure the integrity and reliability of the content made available through its services.

- **Collaboration with Princeton University**: Microsoft has collaborated with Princeton University to create a hub for researchers to access social media data, aiming to enhance the identification and tracking of information operations, with global availability including Europe.
- **Research and Public Data Access**: Bing offers APIs and datasets for public use and research, with no formal requests for data access under DSA Article 40 noted but is working with PEREN in France on a data access request.
- **Other Resources for Researchers**: Microsoft provides a variety of datasets and tools for researchers, including the MS MARCO and ORCAS datasets, and promotes responsible AI practices with tools like the mitigations library.

In-Product Indicators

Below is a summary of Bing's approach to In-Product Indicators which are designed to, where appropriate, create in-product indicators of enforcement actions taken, including broad public notice (e.g., icon noting removed content providing certain details), and updates to users who reported violating content and access to remedies.

Bing adds information to help ensure users are well-informed and not misled or harmed by materials appearing in search results, such as answers pointing users to authoritative sources when standard search results are likely to lack high authority content, adding PSAs or warnings, including a footer to indicate when content has been removed due to legal or policy concerns, and preventing autosuggestions likely to lead to harmful discriminatory materials. Bing works to ensure such materials are available across markets and languages in which Bing is offered. Bing also works to ensure in-product disclosures and disclaimers around engaging with an AI system are written to be consumable by any age.

- **Directing to authoritative sources**: Bing Search provides answers and PSAs directing users to authoritative information on various topics, including public health and safety issues. Further, Bing works to be transparent about how its safety systems work for generative AI features, and for novel systems such as Copilot in Bing reminds users that they are engaging with an AI system, and to remind users of the importance of verifying facts in the source material. Bing ensures that users can find third-party links to content responsive to their queries in standard search results.
- **Information hubs for crises**: During events like public health crises, Bing creates specialized hubs to centralize high authority information, exemplified by the COVID-19 Information hub.
- **Page Insights feature**: The Page Insights feature aids users by providing additional authoritative site information from third-party sources like Wikipedia and warnings or PSAs on sites where there are indications that a site may contain harmful content such as illegal pharmaceuticals or malware.
- **Fact-checking in search results**: Bing's Fact Check feature warns users about false or unfounded content by using third-party fact-checkers who follow Schema.org's ClaimReview protocol.
- **Warnings:** Bing may add warnings (such as for illegal pharmaceutical websites), PSAs, or other authoritative materials to avoid misleading users when harmful content appears in search results.
- **Disclosures:** For Bing conversational features, Bing has implemented in-product disclosures that make it clear to users that they are engaging with an AI product, and link to further FAQs, and explanations about how these features work. Bing's in-product disclosures are written at a 5th grade level to ensure understandability by teen users.
- **Content Integrity:** Bing has also provided users with Content Integrity as a Service enables users to digitally sign and authenticate media using the Fou's digital credentials, a set of metadata that encode details about the content's provenance using cryptography.

Appendix III: Abbreviated Terms

Below are definitions for the acronyms and initialisms mentioned throughout the report.

Acronym	Definition
1CS	One Compliance Assessment, a platform developed by Microsoft to provide a unified compliance system for engineering groups across the organization.
ADL	Anti-Defamation League, a global organization that combats antisemitism, extremism, and bigotry through education, advocacy, and innovation.
AISIC	AI Safety institute Consortium, an initiative by NIST to create safe and trustworthy artificial intelligence (AI).
API	Application Programming Interfaces, a set of rules and protocols that allows different software applications to communicate with each other.
C2PA	Coalition for Content Provenance and Authenticity, an initiative co-founded by Microsoft along with other tech and media companies to develop technical standards for certifying the source and history (or provenance) of media content.
CMP	Content Moderation Platform, an internal tool for implementing content moderation decisions, such as URL exclusion and query treatment.
COPD	Code of Practice on Disinformation, a set of guidelines and standards aimed at combating disinformation and ensuring transparency and accountability in digital platforms.
CSEAI	Child Sexual Exploitation and Abuse Imagery, refers to visual media that contains or promotes child sexual exploitation or abuse.
CSS	Customer Service & Support, internal team that is responsible for providing technical support and assistance to customers.
DDR	Defensive Defect Rate, primary means to measure the efficacy of implemented safety mitigations and the presence of content that violates Bing policies.
DIRT	Defensive Intelligence Response and Transparency, internal team that is accountable for search principles creation, evaluation, and operations.
DPIA	Data Protection Impact Assessment, required for any new product or feature to investigate possible harms to data subjects.
DSA	Digital Services Act, Article 42.4 (a) and (b) of regulation (EU) 2022/2026
DSB	Deployment Safety Board, internal group that reviews impact assessments, including red team testing results and related metrics, prior to product/feature launch for compliance with Microsoft's Responsible AI (RAI) Standard.
DTSP	Digital Trust & Safety Partnership, an initiative focused on promoting a safer and more trustworthy internet by bringing together leading technology companies to develop industry best practices and assessments for trust and safety in digital services.
EAP	Employee Assistance Program, a program that offers counseling and work/life resources to employees and their household adult family members.

EDMO	European Digital Media Observatory, external source Bing leverages for local intelligence from various EU countries.
ERCHT	Enhancing Resilience and Countering Hybrid Threats, a working party that focuses on countering hybrid threats, strengthening the resilience of states and societies to such threats, improving strategic communication, and tackling disinformation.
GDI	Global Disinformation Index, one of Bing's partnerships aimed to empower Bing product teams to take additional actions to promote more authoritative information.
GDPR	General Data Protection Regulation, a comprehensive data privacy law enacted by the European Union (EU) to enhance individuals' control over their personal data and simplify the regulatory environment for international business by unifying data protection regulations within the EU.
GIFCT	Global Internet Forum to Counter Terrorism, organization, in which Bing has a leadership role, dedicated on combating terrorist, violent, and extremist content.
GNI	Global Network Initiative, a multistakeholder collaboration to protect freedom of expression and privacy in tech.
IFES	International Foundation for Electoral Systems, an organization focused on strengthening the cybersecurity practices of investigative journalists who are reporting on abuse of state resources in elections.
IGF	Internet Governance Forum, a multistakeholder governance group for policy dialogue on issues of Internet governance.
IIB	Independent Intermediary Body to Support Research on Digital Platforms, an initiative designed to support independent research on the impacts of digital platforms on society.
INCB	United Nations International Narcotics Control Board, an independent and quasi-judicial monitoring body for the production and trade of narcotics and psychotropics, and their availability for medical and scientific purposes.
IP	Intellectual Property, creations of the mind, such as inventions, literary and artistic works, designs, symbols, names, and images used in commerce.
ISD	Institute for Strategic Dialogue, an independent, non-profit organization dedicated to safeguarding human rights and reversing the rising tide of polarization, extremism, and disinformation worldwide.
IWF	Internet Watch Foundation, a global registered charity based in Cambridge, England remit to minimize the availability of online sexual abuse content, specifically child sexual abuse images and videos hosted anywhere in the world and non-photographic child sexual abuse images hosted in the UK.
KPI	Key Performance Indicators, a measurable value that shows how effectively an individual, team, or organization is achieving key business objectives.
MAU	Monthly Active User, a key performance indicator (KPI) that measures the number of unique users who engage with a product or service within a 30-day period.
MDSS	Microsoft Digital Safety Standard, a comprehensive framework that addresses the means and methods by which online services govern how end users interact with each other in their content and conduct. It provides a set of requirements across Microsoft services and features to ensure digital safety and limit the usage of services for harmful purposes.
MPI	Minutes Per Incident, a metric used to measure the amount of time spent on handling a single incident or case.

MSA	Microsoft Services Agreement, Bing's general terms and conditions governing user behavior.
MTAC	Microsoft Threat Analysis Center, a specialized team within Microsoft that focuses on identifying, assessing, and disrupting threats to Microsoft, its customers, and democracies worldwide.
NCII	Nonconsensual Intimate Imagery, also known as "revenge porn."
NCMEC	National Center of Missing and Exploited Children, a private, non-profit organization whose mission is to help find missing children, reduce child sexual exploitation, and prevent child victimization.
NDI	National Democratic Institute, an organization aimed at strengthening the cybersecurity support infrastructure for political parties and campaigns internationally.
NGO	Non-Governmental Organizations, a group that functions independently of any government with the objective of improving social conditions.
NIST	National Institute of Standards and Technology, a non-regulatory agency aimed to promote American innovation and industrial competitiveness by advancing measurement science, standards, and technology.
OBR	Overblocking Rate, a metric used to measure the rate at which content is incorrectly blocked by a system.
OFCOM	Office of Communications, regulatory and competition authority for the broadcasting, telecommunications, and postal industries in the United Kingdom.
OKR	Objectives and Key Results, a goal-setting framework used by organizations to define and track objectives and their outcomes.
PAI	Partnership on AI, a non-profit organization that brings together a diverse group of stakeholders, including academic, civil society, industry, and media organizations, to create solutions that ensure artificial intelligence (AI) advances positive outcomes for people and society.
PSA	Public Service Announcement, a message in the public interest disseminated without charge, with the objective of raising awareness, changing public attitudes, and behavior towards a social issue.
QC	Quality and Credibility score, a metric used by Bing to evaluate and rank websites. This score helps ensure that users are not misled by problematic material in search results.
RAI	Responsible AI, Microsoft's commitment to cutting-edge research, best-of-breed engineering systems, maintaining a consistently high bar on the policies that Microsoft formulates and promotes, and pursuing excellence with the corporate processes and programs that Microsoft adopts and shares with partners.
RLHF	Reinforcement Learning from Human Feedback, a technique used to improve the behavior of language models by incorporating human preferences into their training process.
SDLC	Software Development Lifecycle, a high-level framework used for developing or changing business applications.
SEO	Search Engine Optimization, the practice of making a public-facing website more visible within search engines, and in turn, to customers and users.
SLA	Service Level Agreement, a contract between a service provider and a customer that defines the types and standards of services to be offered.
SNOW	ServiceNow, a ticket management tool.

SSPA	Supplier Security and Privacy Assurance, a program that holds vendors to a high level of accountability.
TTA	Time to Acknowledgement, a metrics that measures the time it takes to acknowledge a notification or incident after it has been reported.
TTR	Time to Resolution, metric that measures the average time it takes to resolve a customer issue or incident from the moment it is reported until it is fully resolved.
TVEC	Terrorist, Violent, and Extremist Content, refers to content and conduct that can be inherently violent, including calls to action for harm and references to past violent acts.
UHRS	Universal Human Relevance System, a crowdsourcing platform that supports data labeling.
UNCRC	United Nations Convention on the Rights of the Child, an international human rights treaty that sets out the civil, political, economic, social, health, and cultural rights of children.
VLOP	Very Large Online Platforms, an online platform that reaches an average of 45 million or more active users per month in the European Union (EU) and must comply with the DSA.
VLOSE	Very Large Online Search Engines, an online search engine that reaches an average of 45 million or more active users per month in the EU and must comply with the DSA.