Bing

Bing Systemic Risk Assessment Report

August 2025

Microsoft

Contents

Executive Summary....4
Introduction....6
Purpose....6
Document Structure....6
Bing's Approach to Protecting Users and their Fundamental Rights Online....8
Bing's Values and Commitments....8
Bing's Approach to User Safety....9
Generative AI Experiences Specific Mitigations....19
How Bing monitors effectiveness of risk mitigation efforts....22
Bing Overview and Risk Profile....23
User Base and Market....23
Considerations for Bing's Risk Profile....23
Core Search Features....24
Generative AI Experiences....28
Ancillary Search Features Risk Profile....35
Changes to Bing Features....37
Systemic Risk Assessment Methodology....39
Methodology Overview....39
Risk Assessment Process....39
Assessment Tooling....42
Risk Assessment Results....43
Overall Results....43
Changes to Bing's Risk Profile....44
Emerging Trends and Shifts in Global Risk Posture....44
Changes in Risk Rating....45
Results per Risk Area....47
Conclusion....89
Conclusion of Assessment....89
Progress on Enhanced Mitigations from Bing's Last Report....91
Looking Ahead to the Coming Year....92
Appendix I: Detailed Systemic Risk Assessment Methodology....94
Risk Areas....94
Risk Factors and Influencers....95
Risk Assessment Inputs....95
Probability....96
Severity....97
Inherent Risk....98

Maturity of Mitigations 99
Residual Risk 101
Sample scoring calculation 102
Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations 103
Product Quality 103
Product Safety 106
Targeted Interventions 112
Content Moderation 115
Incident Response 119
Research and Partnership 121
Transparency and Reporting 123
Metrics x Funnel 125
Appendix III: Terms and Acronyms 128
Appendix IV: Summary of External Consultations and Partner Insights 131

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, including 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 users and to a healthy society, and Bing must balance these interests and rights to maximize benefits 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 its ability to access answers to the questions they ask.

Bing has implemented a robust and multi-layered approach of five functions and related measures to manage risks and maintain this careful balance. This approach and the implemented mitigation measures 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.

The measures address systemic risk in various ways, including but not limited to: algorithmically ranking content, grounding results in high authority content, removing content when necessary, managing incident response, and enabling user feedback and reports, in order to provide a safe experience for all users. 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.

Bing is also continuously innovating and enhancing its service to help users find exactly what they're looking for faster. Bing's innovation is rooted in its values and respect for user fundamental rights and safety, and Bing considers – throughout the innovation process – the potential for product changes to reduce or increase the systemic risk profile of the service.

Finally, Bing as a service does not exist in a vacuum but within an ecosystem of geopolitical and cultural trends that influence risk vectors. Bing tracks these trends as part of its continuous monitoring to align resources as needed to dynamically address risks. For example, there has been growing concern about the risks posed by GenAI, particularly regarding Non-Consensual Intimate Imagery ("NCII"), Concern about violent extremism also remains high, although with changes in specific actors and targets. Bing uses information on trends to inform product and policy priorities and takes targeted actions consistent with the principles above and described further at How Bing delivers search results.

Continuous monitoring of the ecosystem as well as Bing usage trends is necessary and key to keeping Bing services safe for all users. Annual assessments are also useful, to reflect on trends over discrete periods of time to evaluate whether adjustments in risk mitigation measures are warranted to address more gradual shifts in risk. Therefore, in addition to continuous monitoring activities, Bing conducts an annual assessment of the effectiveness, proportionality, and reasonableness of the implemented mitigation measures 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 ("DSA").

The results of Bing's continuous monitoring and annual Systemic Risk Assessment, factoring in external shifts and internal product changes, indicate that residual risks to users, stemming from use, misuse, or functioning of Bing services, remains consistent and primarily falls within the Low range, with a few exceptions. Bing assesses that the mitigation measures in place across all risk areas are proportionate, reasonable, and effective to reduce risks on the service to an acceptable level, based on implementation of best available scientific and technical insights and the incorporation of internal and external data and expert opinion.

In the 2025 Systemic Risk Assessment, Bing assessed eight of the 12 risk areas as Low Residual Risk. Of the four remaining risk areas, Freedom of Expression and Information is rated as Minimal Residual Risk, while Human Dignity, Illegal Content and Activities, and Private and Family Life are rated as Moderate Residual Risk. The Illegal Content and Activities Residual Risk Rating increased to Moderate solely due to the clerical internal reassignment of risks related to Child Sexual and Abuse Material ("CSAM") from the Rights and Protection of Minors risk area to the Illegal Content and Activities risk area. Based on Bing's unique risk profile as a search engine, the residual risks for the Human Dignity, Illegal Content and Activities, and Private and Family Life areas are within an acceptable range; at the same time, these challenges are evolving and require ongoing monitoring and investment.

Despite the relatively low risk overall, Bing is committed to continually enhancing user trust and safety and, therefore, as part of both this and last year's processes, identified opportunities to further enhance the systemic risk mitigation measures currently in place.

Bing drove progress in each of the six areas highlighted in the 2024 Systemic Risk Assessment report; for example, Bing launched a dashboard showing its core safety metric, the Defensive Defect Rate, aligned to DSA risk areas, for more tailored analysis, monitoring, and reporting. As part of this 2025 Systemic Risk Assessment, Bing has identified the following five opportunities to enhance user trust and safety:

1) Enhancing governance and pre-launch review.
2) Developing and refining policies to align user experience with Bing's Trustworthy Search Principles.
3) Continuing investments in targeted interventions to better manage systemic risks.
4) Enabling greater data-driven improvement by refining documentation, resources, and metrics.
5) Improving children's safety.

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

Purpose

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; ancillary services, including search verticals such as images, shopping, maps, news, travel, and real estate, and advertisements in Bing; and the GenAI-enhanced experiences, Copilot Search, Bing Image Creator, and Bing Video Creator.

The report describes the assessment of systemic risks on the Bing VLOSE impacting the EU for the assessment period of August 2024-July 2025.

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.

Document Structure

This document is structured to provide a comprehensive overview of Bing's user protection measures, its services, and its approach to systemic risk management under the DSA. It begins with a description of the safeguards in place to uphold users' fundamental rights online, followed by a description of Bing, its key features, and its risk profile. The document then outlines the methodology used to identify and evaluate systemic risks across the Bing service. This includes an assessment of how such risks may manifest in practice and how existing mitigations address those risks. Finally, it details Bing's strategy for managing Residual Risks, including prioritization of mitigations based on the risk evaluation outcomes.

The structure is designed to first establish an understanding of the service and its safety framework, and then to systematically examine the alignment between identified risks and the effectiveness of protections intended to mitigate potential harm to users and fundamental rights.

This Report includes the following components:

- **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 coupled with its approach to protecting fundamental rights and (2) Bing's approach to user safety through a multi-layered approach.
- **Bing Overview and Risk Profile:** An overview of (1) the Bing service and its key features across three categories (core search, GenAI experiences, and ancillary search features) as well as updates to Bing features during the assessment period and (2) Bing's risk profile and key considerations due to its nature as a search engine.

- **Systemic Risk Assessment Methodology:** An overview of the process Bing used to conduct this year's Systemic Risk Assessment including an overview of the methodology used, a description of the risk assessment process, and an explanation of the assessment tools used. Further detail, including risk scoring definitions, criteria, and rating scales, is included in Appendix I: Detailed Risk Assessment Methodology.
- **Risk Assessment Results:** A description of (1) each risk area, the risk analysis conducted, and the key relevant implemented mitigations, (2) an explanation of changes to Bing's risk profile, (3) emerging trends and shifts in global risk program, (4) changes in risk ratings, and (5) a deep dive into the results per risk area.
- **Conclusion:** A conclusion of the assessment, a description of progress made on enhanced mitigations from Bing's last report, and a description of the heightened focus areas for the next assessment period.
- **Appendix I - Detailed Systemic Risk 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 – Bing's Digital Safety Risk and Compliance Framework and Mitigations:** An overview of the mitigations that apply across the Bing service and across DSA systemic risks.
- **Appendix III – Terms and Acronyms:** Definitions of the acronyms and initialisms mentioned throughout the report.
- **Appendix IV – Summary of External Consultations and Partner Insights:** A synopsis of Bing's engagement with external parties.

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 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 search and receive 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. These principles also underpin Bing's GenAI experiences, which represents a very small portion of users' experiences. Bing is evolving its policies, operations, and infrastructure to better identify and manage the associated GenAI risks.

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 (further detail on high quality, authoritative ranking under Product Quality).
- 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 SafeSearch.
- 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 provides information about content moderation actionstaken in DSA transparency reports.

AI systems in Bing are developed and evaluated in accordance with Microsoft's Responsible AI ("RAI") 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?

Additional information is available at How Bing Delivers Search Results and in Bing's Webmaster Guidelines.

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

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

More information on Microsoft's broader commitment to responsible practices, including sustainability, ethics, and human rights, is available on Microsoft's Corporate Social Responsibility Report.

Bing's Approach to User Safety

Microsoft respects freedom of expression and the right to access information. At the same time, in accordance with Microsoft principles, policies, and standards for RAI, privacy, digital safety, information integrity and the Bing Trustworthy Search Principles, Bing has implemented a robust and multi-layered system of mitigations to manage risks and maintain this careful balance. These measures are organized into five layers and represented in the Bing Safety Funnel (Figure 1)

The Bing Safety Funnel is applied globally and progresses through five layers:

1. Product Quality;

2. Product Safety;
3. Targeted Interventions;
4. Content Moderation; and
5. Incident Response.

Each layer incorporates specific standards, processes, and mitigations that apply across the Bing ecosystem. As illustrated by the funnel shape, interventions are applied to an increasingly narrower set of queries with highly specific intent. This layered system provides a structured defense that helps users safely find the relevant and trustworthy information they are seeking.

Underpinning the Bing Safety Funnel is Bing's strong emphasis on data-driven safety governance. Specifically, Bing relies on metrics to measure the real-world performance of its safety systems and to guide ongoing improvements. Central to this approach is the Defensive Defect Rate ("DDR"), Bing's core safety metric, which is used to measure the presence of content that violates Bing policies and the efficacy of implemented safety. DDR allows Bing to evaluate the effectiveness of its mitigations, identify regressions, and take corrective action. In addition to overall DDR, Bing now also tracks DDR in the DSA systemic risk areas, enabling more tailored monitoring, analysis, and reporting. Bing reviews DDR regularly in both pre-launch assessments and in production settings. This supports accountability and drives system-wide refinement.

A visualization of the Bing Safety Funnel, with its five layers of safety mitigations working together as a system, and underpinned with metrics, is provided below. An inventory of Bing's service-wide mitigations, organized by the layers of the Bing Safety Funnel, is included in Appendix II: Bing's Digital Safety Risk and Compliance Framework Mitigations.

Figure 1: Bing Safety Funnel



The following sections first highlight *select* service-wide mitigations within each layer of the Bing Safety Funnel to illustrate the broad purpose of the layer and then include the *complete* set of GenAI-specific mitigations applied across the service; mitigation titles are in bold and italic font. Then, they explain how Bing

continuously measures, monitors the effectiveness of this layered approach to safety, and uses metrics to drive continuous improvement. Bing's service-wide mitigations, found in their entirety in Appendix II, broadly apply to all products and features, with specific feature-level mitigations applied in addition. This layered system, based on and continuously monitored through metrics, enables Bing to proactively manage risks and respond to real-world use cases.

Product Quality

At the top of the Bing Safety Funnel is the Product Quality layer. This is where Bing's foundation for safe and relevant search experiences begins. As part of product quality, Bing focuses on engineering and algorithmic design to ensure that users receive high-quality results that align with their search intent while limiting exposure to harmful or low-quality content.

The user's search query is the primary factor determining what information is displayed to users on Bing. The service prioritizes delivering content that is closely aligned with those queries, ensuring users receive information tailored to their individual searches. Unlike platforms that broadly recommend content based on user behavior across the service, Bing's approach centers on query-driven relevance.

When Bing surfaces suggestions, such as autocomplete, trending searches, or homepage modules; these are carefully filtered through targeted mitigations and quality thresholds. This ensures that safety standards are upheld and that harmful or inappropriate content is not elevated. This commitment to precision, relevance, and safety reflects Bing's dedication to maintaining high product quality and providing authoritative information aligned with user intent.

Within Product Quality, Bing's algorithms are the first and strongest defense in delivering safe, authoritative content. These systems are designed to match user queries with the most relevant third-party web pages in Bing's index, elevating trusted sources and reducing visibility of unreliable or manipulated content. Rather than removing content outright, Bing relies ***on continuous improvement of ranking algorithms*** to ensure that harmful content is not surfaced prominently, particularly when users have not explicitly searched for it.

To support this, Bing invests continuously in refining how its crawlers and algorithms detect quality content. These improvements are applied across all markets and languages where Bing operates. Teams regularly assess performance through key metrics, including internal safety indicators, and use these results to inform updates. In addition to technical signals, Bing incorporates feedback from users, Microsoft's policy teams, regulators, and civil society organizations to ensure its ranking decisions reflect evolving standards of trust and accuracy.

Bing's commitment to Trustworthy Search Principles means ***prioritizing high-authority content*** in search results, especially in areas where inaccurate or manipulated content may pose risks. Rather than removing content outright, Bing relies on ranking algorithms to ensure that harmful content is not surfaced prominently, particularly when users have not explicitly searched for it.

To support this, Bing uses a Quality and Credibility ("QC") score, which includes external signals that help identify authoritative sources. These signals help elevate trusted results while demoting pages that may be misleading or exploitative. Bing also uses ranking systems to defend against manipulation, including spam or search engine optimization abuse that could otherwise distort results.

These ranking systems are continuously measured and refined to ensure they reflect user expectations, policy goals, and emerging risks. The result is a more accurate and safer search experience that remains grounded in user intent.

Respecting ***user privacy choices*** is a core Microsoft commitment, and as part of Product Quality, Bing provides users with control over how their data is collected, used, and personalized. Through the Microsoft Privacy Dashboard, users can view, delete, or export their personal data, including search history. They can also adjust advertising and content personalization settings. For example, if a user frequently searches for travel content, Bing might personalize their Microsoft News feed with travel-related stories. These personalization features are optional and can be disabled. In the EU, Bing also requires users to opt in before using cookies or similar technologies that collect data for personalization.

Bing also offers users direct control over what they see through ***SafeSearch***, a feature that uses machine learning to automatically detect and filter adult and explicit content from search results. This feature is especially important in educational and household settings where additional safeguards may be required.

SafeSearch settings can be locked to prevent tampering and ensure consistent application of filters. Bing also provides a mechanism for users to report inappropriate content or suggestions. These reports feed into the continuous improvement of the underlying safety systems, helping Bing respond quickly to emerging issues.

Bing publishes and enforces its content standards through ***terms and conditions***, specifically through the Microsoft Services Agreement and the Bing Image Creator Terms of Use which, among other conditions, prohibit the generation or sharing of inappropriate or harmful content. These policies apply globally and are available in local languages. Bing reserves the right to suspend or ban users who attempt to misuse the service. For example, entering prompts designed to bypass safety systems is not permitted. These terms help ensure that Bing remains a safe and transparent service for all users.

Product Safety

The second layer of the Bing Safety Funnel is Product Safety; a focused set of engineering, policy, and testing mitigations designed to proactively identify and mitigate safety risks. Product Safety mitigations do not focus on content removal. Instead, they help users find high quality information for their query efficiently, for example by surfacing trusted information, adding context, or reducing the visibility of potentially harmful results. Bing's approach is not only deliberate, but also agile, ensuring it can respond quickly and effectively to emerging safety challenges across the service.

Engineering standards and processes establish robust technical foundations to assess and manage a variety of risks. These include rigorous testing protocols, expert reviews, launch readiness assessments, and in-product transparency tools that ensure safe and trustworthy user experiences.

Specifically, before any new feature or product is introduced to users, Bing requires it to pass through a robust ***pre-launch risk assessment***. This assessment process is part of Microsoft's broader compliance framework and includes reviews of privacy, security, accessibility, safety, and responsible AI concerns.

These reviews are conducted by professional compliance managers who ensure that products meet Microsoft's standards and legal obligations. Where risks are identified, teams must implement appropriate mitigations before the product or feature can go live. This process ensures that safety, privacy, and integrity are designed into Bing experiences from the start.

Bing also requires extensive ***red team testing*** to explore safety risks under pressure. These tests simulate how systems might be misused or manipulated in the real world. Multidisciplinary teams and, at times, invited external experts, rigorously test how Bing Search, Copilot Search, and Bing Image and Video Creator respond to adversarial inputs or edge cases.

Red team findings are used to refine training data, harden safeguards, and improve system behavior. Bing ensures that these tests continue even after launch, maintaining vigilance as systems evolve.

Every change to Bing, whether in core search or newer features like Copilot Search, must also pass a ***"shiproom" review***. This structured launch-readiness evaluation ensures proposed updates meet rigorous standards across safety, privacy, security, and accessibility.

Shiproom reviews require product teams to demonstrate the impact of proposed changes on Bing's core safety metrics, such as DDR. Teams must provide evidence of both offline and live (A/B) testing and explain how they've addressed any regressions or new risks uncovered during testing. Further, in the monthly shiprooms, issues and risks are presented to Bing leadership, enabling leadership ongoing governance and oversight. These checkpoints ensure that safety remains a central consideration in how Bing evolves.

Finally, Bing recognizes that safety includes not only what users are exposed to, but also how well they can interpret what they see. By providing contextual information, highlighting trusted sources, and promoting awareness of potential risks, Bing supports users in making informed decisions.

To this end, Microsoft promotes ***digital literacy*** across its services, including Bing. This work is essential in helping users critically assess online content, especially in the face of misinformation or potentially deceptive material, and complements Bing's other proactive safety efforts.

For certain high-risk query categories, Bing displays ***PSAs*** at the top of search results. PSAs highlight authoritative information including reporting and support options, and/or offer clear warnings about potential risks. There are, for example, PSAs in the EU related to the illegal nature of child sexual abuse material and resources and support for queries related to suicide & self-harm. PSAs are one of Bing's most visible tools for helping users navigate sensitive or dangerous topics with clarity and care.

These messages are localized, translated, and adapted to serve users in specific markets where Bing operates. Whether it's a PSA or a warning label, the goal is the same: to provide clear, immediate context when users encounter results that may carry risk.

Bing's safety philosophy includes keeping users informed when safety-related actions are taken. This includes ***in-product indicators and disclosures*** such as warnings, disclosure footers, or icons that explain when content has been removed or flagged, and why. For example, Bing may include a notification in search results when content has been removed due to legal or policy requirements. In AI-powered experiences, Bing includes clear disclaimers that the user is interacting with an AI system, along with reminders to verify content using authoritative sources. These disclosures are written in plain language and designed to be understandable even by younger or less technical users.

Together, these measures, and the full set of Product Safety mitigations in Appendix II, are thoughtful, precise, and continuously evolving. Consistent with the unique risk profile of a search engine, Bing's approach to product safety focuses on measured mitigations, transparency, and ongoing monitoring, rather than removing content or restricting access to information, to support a safe and empowering experience for every user.

Targeted Interventions

The third layer of Bing's Safety Funnel is Targeted Interventions. Targeted Interventions are a combination of algorithmic and manual intervention used by Bing when search results may not align with its Trustworthy Search Principles. These interventions are carefully designed and precisely deployed to address intentional attempts to manipulate search results or to respond to emerging threats in a way that protects users while preserving access to information.

As a search engine, Bing intervenes in limited ways, consistent with Bing's Trustworthy Search Principles to help users find relevant, credible, and safe information while mitigating the potential of unexpectedly encountering results that are harmful, offensive, or illegal.

As part of the ***Safety Measures – Defensive Search Interventions*** mitigation, Bing has a dedicated team accountable for implementing algorithmic defensive search interventions (as well as metrics monitoring and remediation) to address high impact issues in search results, such as information manipulation, hate speech, and other problematic content that could negatively impact user safety.

Bing uses query classifiers to identify queries that may be associated with elevated risk, such as those that could surface degrading, harmful, or exploitative content, and those trigger a set of defensive search interventions. These algorithmic measures are integrated rather than standalone and work together to uphold Bing's relevance, quality, and credibility standards—particularly in cases where harmful content may exploit data voids or manipulate user intent.

When a query is flagged as potentially risky, Bing may respond with measures such as boosting high-authority sources or applying algorithmic dampening to deprioritize low-quality or harmful content. These interventions are applied in accordance with Bing's core search principles and are actively monitored by the dedicated team that manages implementation and assesses impact across high-risk areas. This combined approach allows Bing to detect potentially harmful search patterns early and respond with targeted mitigations that preserve both user safety and access to high-quality information.

In cases where Bing detects threats to search quality in fast-moving or sensitive areas, Bing will ***direct users to authoritative sources.*** This includes areas such as elections, public health, or active manipulation of the information environment. Bing may demote low-authority content, boost more credible sources, restrict problematic autocomplete suggestions, or make manual adjustments to search results to limit harm. Each step is narrowly applied, with the goal of improving user safety while preserving open access to information.

To support this agile defensive work, Bing has a ***Threat Intelligence*** team that specializes in identifying and responding to high-risk search topics. This team continuously investigates where and how search results might be falling short and applies corrections with precision. The team toolkit includes boosting the visibility of authoritative websites, demoting those with low reliability, and restricting suggestions that could lead users toward harmful or misleading content.

Bing treats safety as a living system; constantly monitored and adjusted in response to new signals. There is ***ongoing monitoring of risks*** and metrics to assess whether interventions are working as intended. Bing uses a combination of internal red teaming, user reports, social listening, and threat intelligence to detect gaps and refine its approach.

These mechanisms allow Bing to catch new trends in attempted manipulation, identify emerging high-risk queries, and improve how its systems adapt to changing threats. If a mitigation unintentionally introduces bias or underperforms in a particular region, the team takes swift action to correct it.

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.

Bing's Targeted Interventions helps mitigate the risk of users expectedly encountering harmful, offensive, or illegal content. The goal is not to censor or overcorrect, but to intervene in targeted ways leveraging a range of resources to support users' freedom to explore, question, and discover.

Content Moderation

The next layer in the Bing Safety Funnel is Content Moderation, which consists of narrowly scoped algorithmic and manual interventions applied when search results conflict with Bing's Trustworthy Search Principles due to legal, regulatory, or policy-based reasons. This layer is used with restraint and precision, focusing only on legal obligation and the clearest and most sensitive areas of harm.

In certain jurisdictions, Bing is legally required to remove access to specific search results. These laws may permit individuals, organizations, or governments to request the removal of indexed pages for reasons such as copyright infringement, libel, defamation, unauthorized publication of personal information, or other content that may violate personal rights or public order. Where such requests are submitted, Microsoft carefully evaluates each one to determine its legal basis, the scope of the request, and the requesting party's authority.

Bing's legal review process aims to strike the right balance between compliance with local laws and the preservation of fundamental rights to freedom of expression and access to information. Legal removal areas where Bing applies a consistent global approach include the removal of child sexual exploitation and abuse imagery ("CSEAI") and enforcement of copyright takedown requests.

Bing maintains rigorous and principled processes, policies, and procedures for ***content removal and other mechanisms to counter harms***. Central to these processes is a firm commitment to the prevention of CSEAI. Bing proactively scans content to detect known CSEAI using hash-matching technologies such as PhotoDNA and MD5. This scanning occurs before content enters the search index and is used to ensure such material does not appear in results at all, across Core Search and Copilot Search. Bing also proactively scans image uploads in Bing Image Creator and Bing Video Creator to prevent known CSEAI from being uploaded. This proactive detection is supplemented by intelligence from trusted partners, including NGOs and law enforcement, and through user reporting channels. When CSEAI is identified, Bing removes it globally and without hesitation. This is a zero-tolerance area where Microsoft's commitment is absolute.

Bing also works to remove adult or sexually explicit images of another person online when shared without that person's consent. This is commonly referred to as NCII, or "revenge porn". Microsoft considers this to be a gross invasion of personal privacy – whether the content is real or a "deepfake" image that is created using AI or other photo editing tools

Finally, Bing also removes certain other content under global legal standards or Microsoft policies, such as materials containing extremely sensitive personal information that could be exploited for fraud or identity theft. Removals may be prompted by legal orders from governments, user-generated reports, or policy-based decisions following internal investigations. Microsoft applies rigorous standards to each request, including the potential impact on user rights and public interest.

Bing provides users with an accessible ***report a concern*** mechanism on every page of Bing Search and across GenAI experiences. Through this function, users can raise concerns about potentially unlawful or harmful content. These reports are routed to human moderation teams who assess submissions based on Microsoft's policies, relevant laws, and the specific context of the reported material. While not all reports lead to content removal—especially in cases involving lawful but offensive or disputed content—each report contributes to Bing's broader understanding of emerging risks and informs future updates to its safety interventions.

Bing's commitment to privacy and dignity is also reflected in its implementation of the ***Right to Be Forgotten ("RTBF")***. Under this policy, users in jurisdictions such as the EU can request the removal of search results associated with their name if those results are outdated, inaccurate, irrelevant, or excessive. Requests are

reviewed by trained staff using clear criteria to balance the individual's right to privacy with the public's right to access information. Where appropriate, Bing removes URLs and associated search suggestions from the index in response to validated requests. Escalation pathways are in place for complex cases, including collaboration with local courts and data protection authorities where necessary.

This process is underpinned by strong training protocols and quality assurance systems. Bing ensures consistency and fairness in RTBF decisions, while maintaining transparency about the process through user guidance and regular engagement with regulatory authorities.

Bing's content moderation policies are shaped by Microsoft's broader commitments to digital safety, freedom of expression, human rights, and compliance with international law. While moderation is applied in limited and clearly defined scenarios, these efforts are critical to protecting users and maintaining the trustworthiness of the search experience.

Incident Response

Incident Response forms the final and most dynamic layer in the Bing Safety Funnel. It is designed to respond swiftly and effectively when urgent quality or safety concerns arise. These manual interventions come into play when internal monitoring, external escalation channels, or user reports identify issues that require immediate attention to uphold Bing's Trustworthy Search Principles.

Unlike earlier layers that rely on ranking systems, safety engineering, or defensive interventions, to guide safe and high-quality experiences, Incident Response serves as Bing's rapid-response capability. It reflects Bing's commitment to thoughtful stewardship of search results; which ensures users can trust that when issues emerge, they are addressed with care, speed, and accountability.

At the core of ***Incident Response and Service Level Agreements (SLAs)*** is a well-established infrastructure for responding to notices involving illegal or policy-violating content. This includes materials such as CSEAI, NCII, exposed personal information, and other high-risk violations. When Bing receives a credible report from governments, users, or trusted third parties, it acts swiftly to evaluate and address the concern.

Bing's support teams are trained in legal and policy standards, with clear procedures for escalating complex cases to local legal experts and subject matter specialists. Reports are tracked through Microsoft's internal incident management system, which ensures accountability and drives collaboration across engineering, legal, safety, and policy teams. Alerts are sent to designated stakeholders, and all incidents are triaged according to priority, with time-sensitive issues handled under strict SLAs that form part of Bing's broader Objectives and Key Results (OKR).

The prioritization process considers the nature of the report, the sensitivity of the policy area, and contextual details such as region, language, and media format. While response times vary based on these factors, most reports involving high-priority concerns are resolved within 24 hours. This fast-paced yet thoughtful process ensures that Bing's response mechanisms remain both effective and user-respecting.

Regular testing and quality checks are in place to ensure that these workflows operate reliably. Bing's commitment to continuous improvement means that incident response systems are refined over time based on lessons learned, stakeholder feedback, and evolving safety challenges.

Because Bing is a search engine, not a content-hosting service, its users do not post content directly to the service. As a result, Bing does not maintain user-facing ***appeals*** processes for content removals. However, where Bing applies enforcement actions, such as when removing or downranking search results in accordance with policy or legal obligations, it offers webmasters a means of appeal.

Through Bing Webmaster Tools, site owners who believe their content has been incorrectly impacted can request a review. These appeals are evaluated by trained reviewers, supported by escalation paths to experts where needed, to ensure outcomes are consistent with Bing's principles and commitments to fairness, transparency, and trust.

In addition to its core search product, Bing also supports services such as advertising that involve direct user participation. In these contexts, Bing enforces clear standards to prevent misuse that could harm users or compromise the integrity of the service. Advertisers who engage in fraudulent, misleading, or illegal behavior are subject to immediate ***suspension of advertiser accounts***.

Bing applies these actions judiciously and in alignment with Microsoft's broader safety commitments. Suspension decisions are made based on evidence and internal enforcement guidelines, with appropriate documentation and oversight to ensure consistency and due diligence.

Cross-Funnel Research, Partnership, Transparency, and Reporting

In addition to the five layered system of mitigations, Bing engages in research and partnership with a variety of internal and external experts to inform its standards, policies, practices, and processes. Bing is also transparent about its practices and assesses and reports on safety through various voluntary and regulatory channels. This is illustrated through the mitigation measures described herein.

Research and Partnership

Bing collaborates closely with ***dedicated cross-functional teams of experts*** across Microsoft. These partners specialize in areas such as digital safety, election integrity, and RAI. They help develop and maintain company-wide policies that guide Bing's decision-making on safety interventions and mitigation strategies. Their input ensures that Bing is aligned with evolving regulatory expectations and public interest standards.

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 information manipulation. Bing is evaluating additional research data sharing opportunities pertaining to elections.

Both Bing and specialized cross-Microsoft teams regularly ***engage externally with 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.

Bing operates a 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 consistent with the recent Delegated Regulation.

Bing regularly ***engages with regulators*** around the world to understand key concerns, share information, and incorporate feedback into product design and safety systems as appropriate.

Furthermore, Microsoft has developed tools and multistakeholder ***third party partnerships*** to combat the rise of misinformation that can pose risks to public health or negatively impact physical or mental wellbeing.

Partnering with independent third-party organizations empowers Bing product teams to take additional actions to promote more authoritative information. Microsoft is also a founding member of the Coalition for Content Provenance and Authenticity ("C2PA") alongside Adobe, Intel, TruePic, Twitter, the BBC, and other tech and media companies. The coalition has an open-source content provenance tool, which allows creators to claim authorship while empowering consumers to make informed decisions about what digital media to trust.

As one example, Microsoft worked with independent third-party partners to perform red team testing of Bing Image and Video Creator to understand the risk of misleading images generated by Bing Image and Video Creator. As part of its analysis, the independent third party prompted Bing Image and Video Creator to create visuals that reinforced or portrayed prominent false narratives related to politics, international affairs, and elections. Based on these evaluations, Microsoft improved existing Bing Image and Video Creator mitigations.

Microsoft, including key members of the Bing Search team, actively participates in the ***Partnership on AI ("PAI")***, collaborating with peers across industry and civil society to identify and mitigate potential harms associated with synthetic media. As part of this work, Microsoft has contributed to the drafting of the PAI's Synthetic Media Code of Conduct; which are guidelines that promote the ethical and responsible creation, distribution, and use of AI-generated content, including deepfakes and digitally altered media.

Transparency and Reporting

Bing practices transparency by notifying users through notices on the search engine results page when content is removed and by publishing ***transparency reports***. 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.

Bing provides ***user transparency*** and 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.

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.

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.

Bing is a signatory to the ***EU Code of Conduct on Disinformation ("COCD"), formerly known as the Code of Practice on Disinformation,*** and subscribes to many commitments in the code dedicated to reducing disinformation risks, including providing users with indicators of content provenance, supporting good faith research into disinformation by enabling researcher access to data, and providing biannual reporting pursuant to the code on systems and policies in place to prevent disinformation on Bing across the EU, among others. Bing also provides enhanced reporting on agreed upon elections and other declared "crisis" under COPD.

As a search engine, Bing is committed to supporting and ***prioritizing users' fundamental rights*** and their rights to access and seek information. Bing does not broadly remove content from its index. Instead, it focuses on minimizing unintended exposure through ranking and targeted filtering.

Generative AI Experiences Specific Mitigations

Given the unique and evolving nature of GenAI capabilities GenAI receives dedicated consideration within Bing's risk management framework. Like other transformational technologies, harnessing the benefits of AI is not risk-free. A core part of Microsoft's RAI program and Standard is the requirement to identify potential risks, measure their propensity to occur, and build mitigations to address them. These risks are addressed through a defined set of mitigations, including:

Product Quality

GenAI Terms and Conditions governing the use of Bing's GenAI features are outlined in the Bing Image Creator Terms of Use. These provisions prohibit activities that could negatively impact users, such as engaging in harmful, fraudulent, or illegal activities, violating privacy, or creating inappropriate content. Users who violate these terms may face limitations on their access to the services.

All Microsoft products that incorporate artificial intelligence, including Copilot Search and Bing Image and Video Creator, are governed by ***Microsoft's RAI Standard***. This standard requires comprehensive reviews and mitigation planning before AI-powered features are launched.

These reviews focus on identifying potential risks, assessing the likelihood and severity of harm, and ensuring mitigations are robust, context-appropriate, and in line with Microsoft's AI Principles. For GenAI experiences, this often includes added layers of technical safeguards, transparency mechanisms, and human oversight. By rigorously evaluating generative features before deployment, Bing prioritizes user safety and dignity, while still enabling innovation. This process reflects Bing's commitment to maintaining a respectful, secure digital environment in the face of emerging AI capabilities.

To guide GenAI models in generating responses that align with Microsoft's AI Principles and user expectations, Bing employs a technique known as "***metaprompting***." Metaprompting involves providing instructions to the AI model to influence its behavior, ensuring that the system operates responsibly and ethically.

Copilot Search responses are generally ***grounded in high-ranking search results***, a process known as "grounding." This approach involves providing data to a large language model (LLM) to improve its ability to generate accurate, relevant, and updated outputs. By centering responses on high-authority content from the web, Bing aims to prevent the generation of content containing private information hosted on low-authority sites. Additionally, providing links to source websites allows users to evaluate the credibility of the information presented.

User-centered design and user experiences are essential aspects 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 Search and Bing Image and Video Creator for the first time, these services offer various touchpoints to help them understand the capabilities of the system, disclose that they are powered by AI, and communicate limitations. Elements of the experience also help users better understand its interactions, including chat suggestions specific to RAI, explanations of limitations, ways to learn more about how the system works, and easily navigable references to show users the results and pages in which responses are grounded.

Copilot Search and Bing Image and Video Creator include ***GenAI in-service disclosures*** and reminders to users that AI can make mistakes, encouraging them to verify the information generated. Furthermore, Bing provides FAQs, help pages, and other public-facing information sources to educate users on the nature of AI-driven search experiences, including their uses, safeguards, and limitations.

Copilot Search responses include ***citation links***, and the corresponding web results are displayed at the bottom of the results. This practice allows users to verify the information and assess the credibility of the sources.

To address issues related to ***conversational drift***, Bing has limited the number of turns per chat session in Copilot Search. This limitation helps prevent responses that are repetitive, unhelpful, or inconsistent with the intended tone. The team continues to evaluate additional approaches to mitigate this issue.

In cases where a user's prompt is ambiguous, Copilot Search and Bing Image and Video Creator may use the LLM to build out more details in the prompt to help ensure users receive the response they are seeking. This ***prompt enrichment*** does not rely on any knowledge of the user or their prior searches but instead on the AI model. These revised queries are visible in the user's chat history and can be deleted using in-product controls.

Bing empowers users to make informed decisions about their ***privacy choices*** and how their data is collected and used in GenAI experiences. Through the Microsoft Privacy Dashboard, authenticated users can view, export, and delete their stored conversation history, including any queries or prompts submitted in Copilot Search.

These controls complement existing privacy settings, such as cookie banners and consent mechanisms, which continue to apply to generative features. Bing also applies an internal review process to examine how choices are presented, ensuring users clearly understand its options and the implications of its decisions. By giving users meaningful control over its data, Bing supports autonomy and accountability in the use of AI-driven technologies.

Product Safety

Transparency is essential in building trust in AI. For ***AI Transparency***, Bing displays prominent disclosures and in-product notices when users are engaging with AI-generated content. For example, in Copilot Search, an information icon near the feedback buttons alerts users that the response was generated by AI. Hovering over this icon reveals a plain-language explanation, with links to more detailed documentation. These disclosures are written at a fifth-grade reading level to ensure they are accessible to young users and audiences of varying digital fluency. To support informed engagement and reduce manipulation, Bing has also introduced content integrity as a Service, a system that uses digital credentials and cryptography to authenticate the origin of media. This helps users trust that the images or media they are seeing have not been tampered with, and that they come from verified sources.

Bing has certain ***age restrictions for GenAI features***. While some, limited functionality in Bing Image and Video Creator is available to all unauthenticated users, full functionality is restricted to authenticated users who meet the appropriate age thresholds. Users must be signed in with a Microsoft account that reflects an age of 13 or older, or a higher age in jurisdictions where parental consent laws apply, to access full functionality, including multiple turns on the same content. If a user's account indicates they are below the required age, they are prevented from accessing Bing Image and Video Creator, even the limited experiences available to unauthenticated users. This age-based access control helps protect younger users from exposure to content or interactions that may not be appropriate for development.

Bing Image and Video Creator uses targeted safeguards to limit the generation of inappropriate or potentially harmful content. A key intervention is the use of intelligent ***blocklists***—filters designed to detect and restrict specific user inputs that contain sensitive or risky names, terms, or phrases. These inputs are sorted into "hard" or "soft" blocklists, depending on the level of sensitivity and risk associated with them.

The development and refinement of these blocklists is a collaborative process involving internal product and safety teams, insights from jailbreak researchers, red teams, and specialized legal counsel. For example, during election periods, Bing consults with elections counsel to ensure the blocklists reflect heightened sensitivities around electoral misinformation or manipulation. This multi-disciplinary approach allows Bing to tailor safeguards to real-world risks in a dynamic and proactive manner.

The development and deployment of GenAI features at Bing are guided by safety processes that ensure adherence to high standards of responsibility and ***emphasize safety by design for GenAI features***. These processes include red team testing, expanded reactive social listening systems, prominent reporting functionality, and operations and incident response mechanisms. These measures help identify vulnerabilities and inform the development of targeted mitigations.

To address risks associated with Bing generating harmful content, Bing has placed specific, ***targeted limitations on GenAI text and image responses*** for certain queries. For example, when Bing classifiers identify potentially harmful content in user prompts for Copilot Search and Bing Image and Video Creator, Bing may not return generated content to users, may divert users to different topics, or redirect users to traditional search results. These measures help prevent the dissemination of harmful or inappropriate content.

Targeted Interventions

Bing deploys specialized mitigations in its GenAI features, such as "***classifiers***," to reduce the risk that harmful content might be generated by Bing. These classifiers analyze text, images, and videos to identify potentially harmful content in search queries and generated responses. By applying AI-based classifiers and content filters, Bing can flag and prevent the production of harmful content, including in features such as Copilot Search and Bing Image and Video Creator. For instance, classifiers have been implemented to mitigate the risks of generated content related to public figures and to prevent the generation of low-authority content.

Content Moderation

Copilot Search uses the same content moderation processes as Core Search, and users are able to report concerns through the same features and tooling. Additionally, Bing allows individuals, such as artists, creators, or brand owners, to request that Bing include their name on the ***Bing Image and Video Creator blocklist*** so that the GenAI models do not replicate their artwork.

Incident Response

In serious cases, Bing may ***restrict a user's access to generative services*** or suspend a user's account when there is evidence of repeated or intentional misuse. This includes attempts to bypass content safety protections or efforts to generate highly sensitive material. User's who demonstrate clear intent to create CSEAI will have their accounts suspended.

For Bing Image and Video Creator, users whose prompts are repeatedly flagged for policy violations may be automatically suspended for a defined period or permanently banned in severe cases. To support fairness and transparency, these users are given an opportunity to appeal such decisions.

Appeals are reviewed by trained human moderators who consider the user's behavior, applicable terms, and context. Based on the outcome, the enforcement decision may be upheld, or the user's access reinstated. This process ensures that even enforcement actions within GenAI experiences are guided by a commitment to procedural fairness and proportionality.

How Bing monitors effectiveness of risk mitigation efforts

As noted above, Bing is committed to data-driven governance and relies on measurements across the Safety Funnel to track and manage online safety risks across its services. Metrics are considered holistically across Bing's mitigations, such as the overall Defensive Defect Rate ("DDR"), Bing's core safety metric and the DDR for each DSA systemic risk area, as well as threat intelligence, defensive search interventions, user reports, social listening, that work together to evaluate the impact of mitigations, surface issues, and guide continuous improvement. Each plays a distinct role in the system.

Measurement metrics

Bing uses DDR as a primary means to measure the service of how often harmful or low-quality content appears in search results by using "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 monitors its overall DDR, as well as DDR based on the DSA systemic harm areas. Bing sets objectives to improve overall DDR 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 enhancements has allowed Bing to keep improving the safety of the service over several years.

In addition to quantitative signal of safety performance of DDR, Bing's commitment to metrics spans its layered system of mitigations, and each has distinct value. For example, metrics related to red teaming and threat intelligence proactively identify new manipulation tactics; defensive search interventions capture the performance of targeted algorithmic responses in high-risk contexts; user reports offer direct feedback from real-world users; social listening captures broader sentiment and trends; and incident reporting provides visibility into the effective of Bing's response processes. This integrated approach, continually monitored and adapted based on new signals, enables Bing to both identify gaps quickly and adapt to emerging threats and assess the longer-term effectiveness of safety mitigations to inform future investment.

Bing Overview and Risk Profile

User Base and Market

Bing has approximately 133 million average Monthly Active Users ("MAU") in the EU during the period of August 2024 to April 2025. 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 computer.

Bing has both authenticated and unauthenticated users. With the exception of optional image and video generation features that require authentication, Bing features, including other generative AI experiences such as summarizing articles or documents, translating, writing, rewriting in different tones or styles, 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.

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. 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 July 2025.

Considerations for Bing's Risk Profile

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 service, data-related practices such as communications, Bing tends to be a lower vector for systemic harms than may be present on other services. Because Bing generally does not allow users to post or share their own content on the service, unlike in social media, there is limited risk related to user behavior toward other users on the service, and limited risk on the service that sharing of 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 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 GenAI features – user behavior on the service 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 GenAI experiences in Bing were designed to provide new ways for users to find the information they want, making it easier for users to find answers to their research questions, and answer more sophisticated questions or inspire new creativity with images. 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 an increased risk that users could take content generated using these tools and share it separately on third party services and 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 GenAI 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 associated with advertisements on the service and works to ensure that the advertisements appearing on the service meet 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.

Core Search Features

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 CSEAI when identified.

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 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 from a variety of third-party sources.

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.

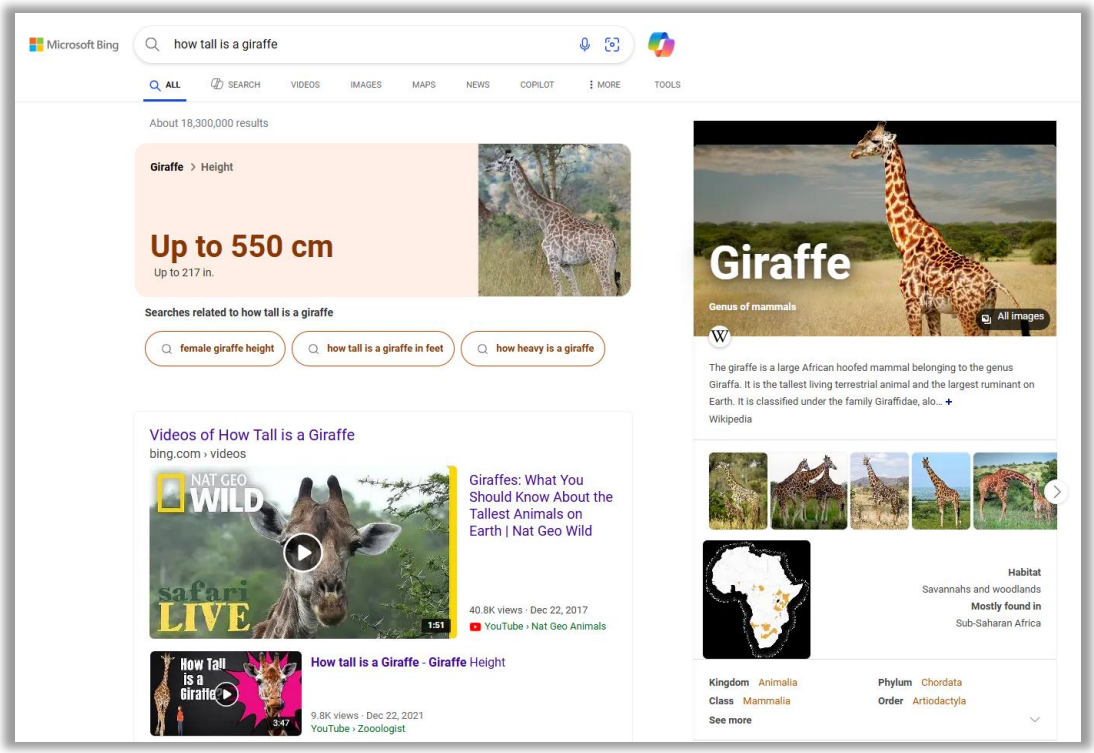
How Search Works

Returning search results involves complex, near-instantaneous algorithmic calculations. The first step in building search is figuring out which pages exist on the Internet 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.

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 550 cm. for males" 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 GenAI to provide an even richer experience, linking to key sources.

Figure 2: Core search results screenshot



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 3: Search suggestions screenshot

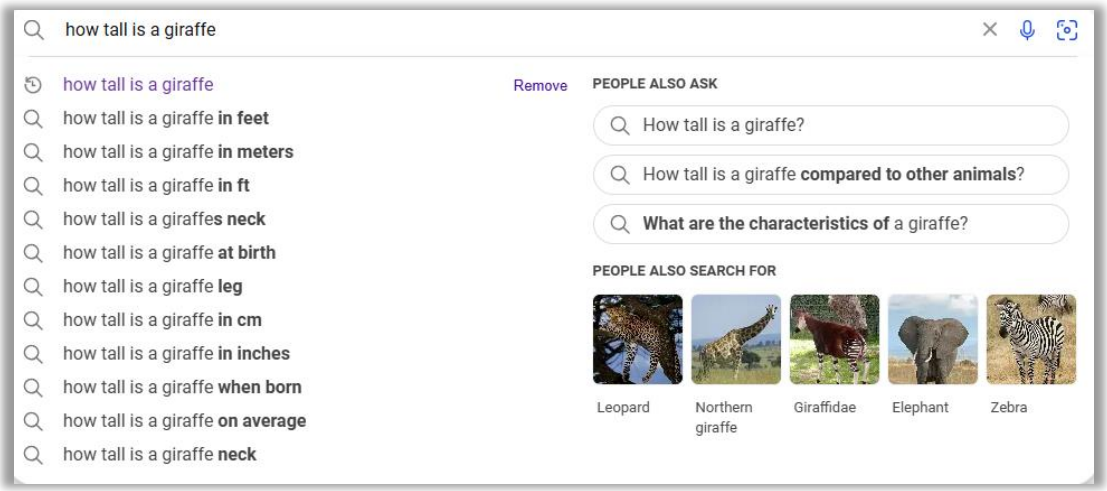
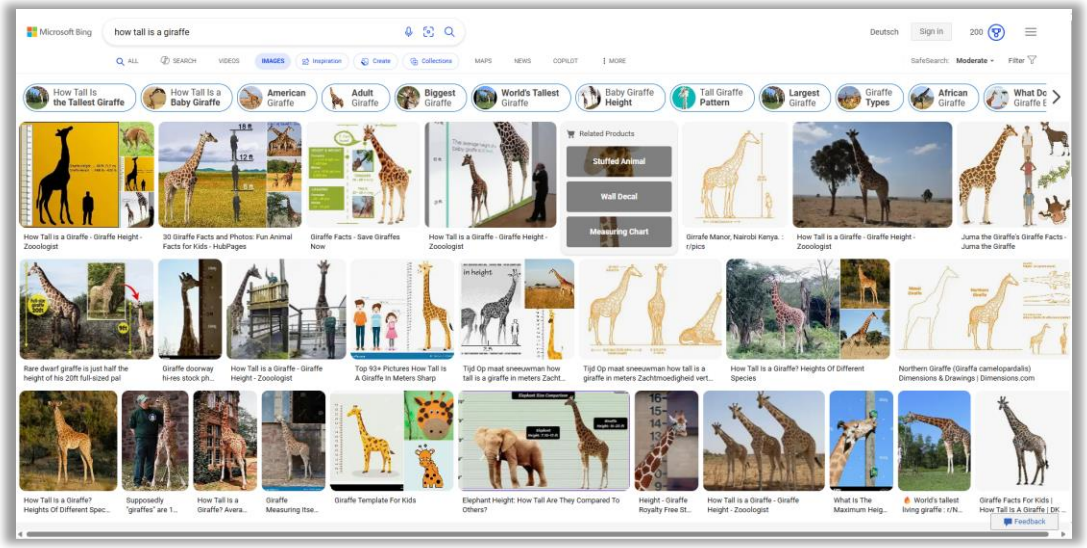


Image and Video

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 4: Bing Image results screenshot

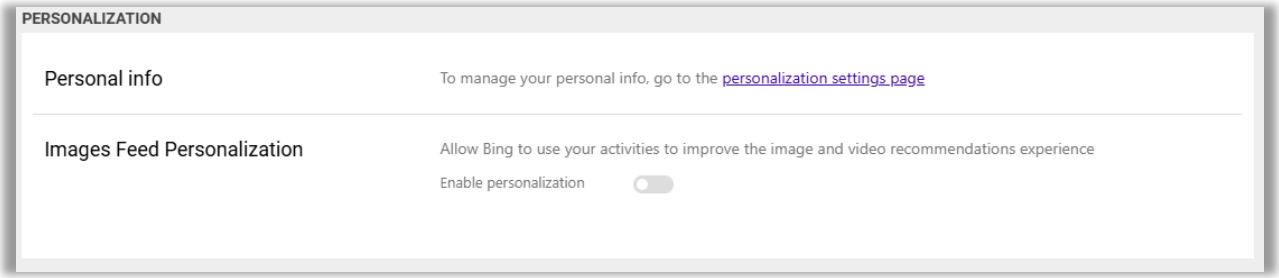


When a user first lands on the image or video experience homepages, prior to entering a query, the homepage may show recommended content based on prior interactions with the site, such as the user's search history, engagement with image results, saved images, and - where a user allows - their Edge browsing 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.

Visual Search

In Visual Search 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 Bing to understand the context, interpret, and answer questions about the image. Users can also upload their files to Bing to interpret, convert, process, or calculate information from them.

Figure 5: Personalization toggle screenshot



Core Search Risk Profile

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.

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.

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

- Risk that bad actors manipulate search results or exploit data voids to surface low authority, low quality content with an intent to deceive or manipulate audiences.
- Risk that search engines return content that provides hateful rhetoric or advocates the use of violence against people because of their race, or other protected traits in the top search results.
- Risk that search over moderates the search results, removing or downranking URLs limiting access to information.
- Risk that search engines return otherwise illegal content in the top search results.
- Risk that search engines return potentially distressing health or other information that can have a negative impact on users' mental health.
- Risk that search results or autosuggest features lead users to content that includes NCII or leads users, including minors, to synthetic NCII content.
- 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 engines return content that is inaccurate, misleading information, harmful misinformation, related to crisis events or issues of public security in the top search results.
- Risk that search engines return content that instructs on, normalizes, promotes, or facilitates sexual exploitation, grooming, or abuse of a child.

The key mitigation measures Bing has implemented to address these potential risk manifestations are described in the Bing's Approach to User Safety section.

Generative AI Experiences

GenAI experiences offered via Bing include Copilot Search, Bing Image Creator, and Bing Video Creator, as well as a variety of embedded experiences such as summarizing articles, translating, writing, and rewriting content in different tones and styles.

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 GenAI tools, including Copilot Search and Bing Image and Video Creator, are operated independently of OpenAI, and Microsoft has added

additional safety mitigations to the underlying models, including combining with search algorithms, adding bespoke metaprompts to guide GenAI 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, and images. Summarizations that rely on web search results 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 on these links to learn more about a topic and the information relied upon in the summary or chat response.

Copilot Search

Copilot Search seamlessly blends the best of traditional and generative search together to help users find what they need and represents a significant advancement in transforming the search experience. Copilot Search is an update to the previously shipped product, Real Time Generative SERP (search engine results page) in Deep Search. Copilot Search, like Real Time Generative SERP in Deep Search before it, responds to the user query with Generative Text. The text is generated by using an LLM to understand the intent of a user’s query, issuing additional search queries to Bing to understand the original user query, and then synthesizing a response which is grounded on the retrieved search results. Real Time Generative SERP and Deep Search no longer exist as they evolved into Copilot Search.

With respect to features and functions, Copilot Search is similar to Real Time Generative SERP in Deep Search experience. Depending on the query, the user will receive an easy-to-digest summary of the most critical points, a clear answer, or a smart layout of information to help them find what they are looking for. Bing has set out to deliver helpful, clearly cited sources, plus rich, relevant data, images and videos from the users’ favorite publishers and content owners to inspire the user to explore further. There is no modification to ranking of web sites in the search results that are used by grounding.

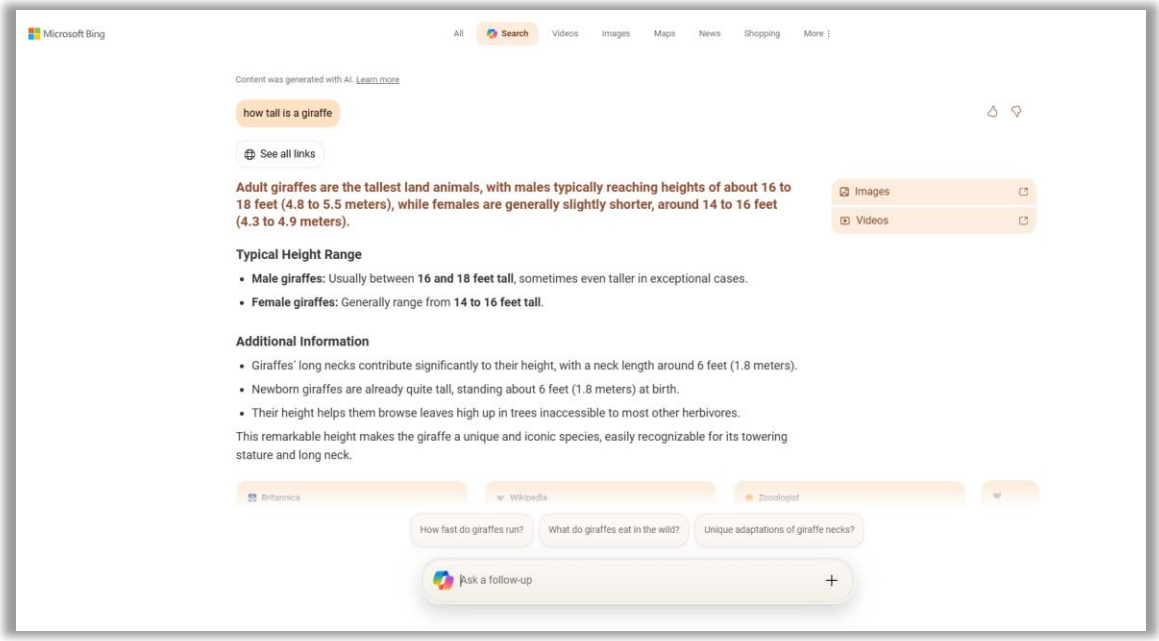
Bing has designed Copilot Search with publishers and content owners in mind to support a healthy web ecosystem. In addition to the citations within the generative responses, Bing inline links the entire sentence or passage within the responses so the user can easily navigate to the sources. Bing also includes easy access to cited sources and all relevant web results at the top and they’re also highlighted prominently at the bottom of the result. This allows the user to be just a click away from their favorite publishers and content owners. Each Copilot Search result will have suggestions on related subjects. These clickable topics help the user uncover new information and make deeper exploration easier than ever – all without losing context in a multi-turn experience. Once a user clicks a topic, results are quickly summarized, and additional follow-up searches will be suggested so the user can continue exploring. Plus, in Copilot Search, the users’ previous result is accessible from the same results page, making it simple to compile information around a single topic in context.

New features that are added to Copilot Search include the ability for the user to ask follow-up questions / queries about the generated response. These follow-up questions can be either free-form user queries or pre-defined suggestions. Users that are not signed in have a 30-follow-up question / query limit per day whereas users that are signed in have a 100-follow-up question / query limit per day. Also, in Copilot Search, depending on the query, there can be different experiences contained within a single page, including but not limited to a GenAI text summary with references, sections of creative text (creative writing, code generation, math), or a creative image, supported by Bing Image or Video Creator. This image creation functionality is the same as what exists on bing.com.

Copilot Search will in some instances, not provide a GenAI experience because responses may be inconsistent with Bing's policies and/or applicable law. In these cases, Copilot Search will return traditional core search results related to the user prompt.

Users can access Copilot Search by going to aka.ms/CopilotSearchinBing and start exploring. Users can also access Copilot Search via bing.com either through the navigation bar or simply by clicking on the suggested, related topics right under relevant answers in Bing.

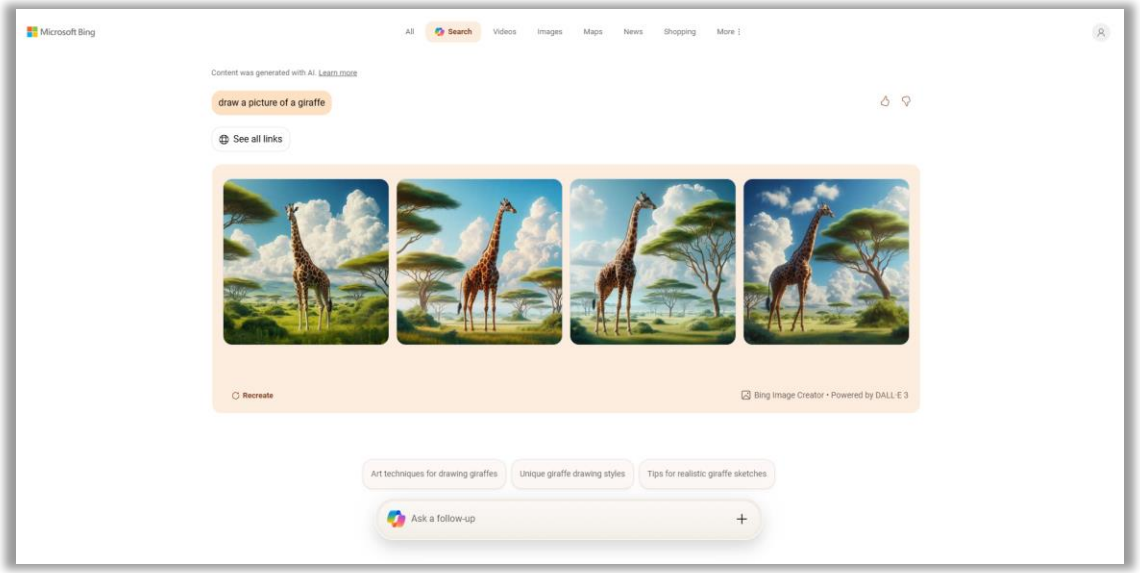
Figure 6: Copilot Search results screenshot



Bing Image Creator

When Bing Image Creator receives a user request to generate an image (for example, the prompt begins with "draw me a …"), the system will, in most cases, generate an image response to the user's prompt. By typing in a description of an image and other context, such as artistic style or key image details desired, Bing Image Creator will generate a response, which is then displayed. For example, a user can type "create an image of a giraffe" as a prompt in Bing Image Creator to generate graphics depicting a giraffe. Bing Image Creator relies upon an image generation model from OpenAI, with additional Microsoft safety measures implemented. With Bing Image Creator, users can not only generate images using prompts but can also resize or restyle them by choosing from different restyle options. Authenticated users signed in with a personal Microsoft account have 15 "boosts" (fast creations) a day, after which image generation will move to a slow queue resulting in longer generation time. Authenticated users signed in with a work or school account are not able to create further images after they use all of their boosts.

Figure 7: Bing Image Creator results screenshot



Bing Video Creator

When Bing Video Creator receives a user request to generate a video the system will, in most cases, generate a visual video responsive to the user's prompt. By typing in a description of a video and other context, such as artistic style or key frame details desired, Bing Video Creator will generate a response, which is then displayed. For example, a user can type in "create a video of a giraffe walking" as a prompt for Bing Video Creator to generate a video. The generated video is a dynamic set of frames that does not have any audio.

Generative AI Experiences Risk Profile

While users can create AI-generated content using GenAI features available on Bing, and users can 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 GenAI 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 GenAI or search users.

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

- Risk that GenAI summaries are based on low authoritative sources and produce content that may be harmful or misleading.
- Risk that adversarial user queries or grounding data may bypass the safety features for GenAI and produce harmful or misleading content.
- Risk that GenAI Protection Systems flag and block the generation of legitimate content.
- Risk that GenAI summaries inaccurately interpret source data to generate incorrect or misleading information about an individual that could be defamatory.
- Risk that oversimplified GenAI summaries generate potentially distressing health or other information that can have a negative impact on users' mental health.
- Risk that GenAI returns content that could be used to find applications, tools, or services for creating synthetic NCII in the top search results.
- Risk that GenAI returns content to minors that normalizes eating disorders in the top search results.

- Risk that bad actors misuse Bing Image or Video Creator through generating images that are used to intimidate, harass, or coerce political or civic actors into complying with their demands.
- Risk that Bing Image or Video Creator is used to produce images or videos that degrade or discriminate against specific groups.
- Risk that Bing Image or Video Creator disproportionately generates images that align with particular cultural standards or viewpoints, resulting in less diverse representations of the population.
- Risk that Bing Image or Video Creator could be misused to generate synthetic, yet realistic private or intimate images of individuals.
- Risk that Bing Image or Video Creator is used to generate images that contribute to public security risks.

The key mitigation measures Bing has implemented to address these potential risk manifestations are described in the Generative AI Experiences Specific Mitigations section.

Ancillary Search Features

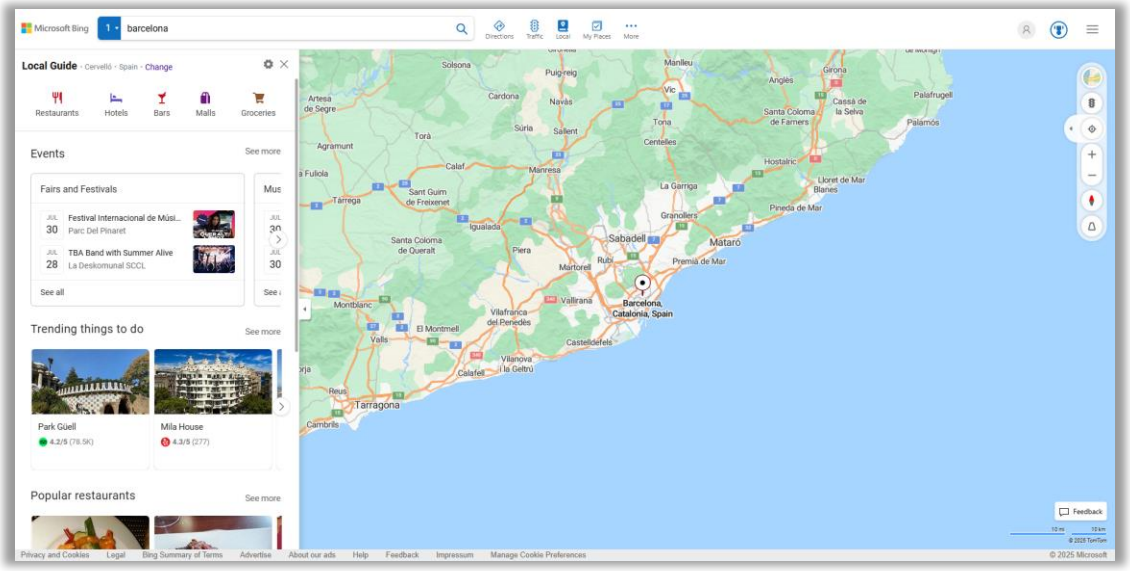
Bing Core Search links to a number of "verticals" or features areas to enable enhanced search functionality, including:

Maps

Bing Maps allows users to search for information about businesses, landmarks, locations, and other geographic data in a map interface. In France and Germany, Bing Maps offers 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.

Maps' Commute feature allows users to identify their home and work to provide a personalized route. With the destinations feature, users can 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 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.

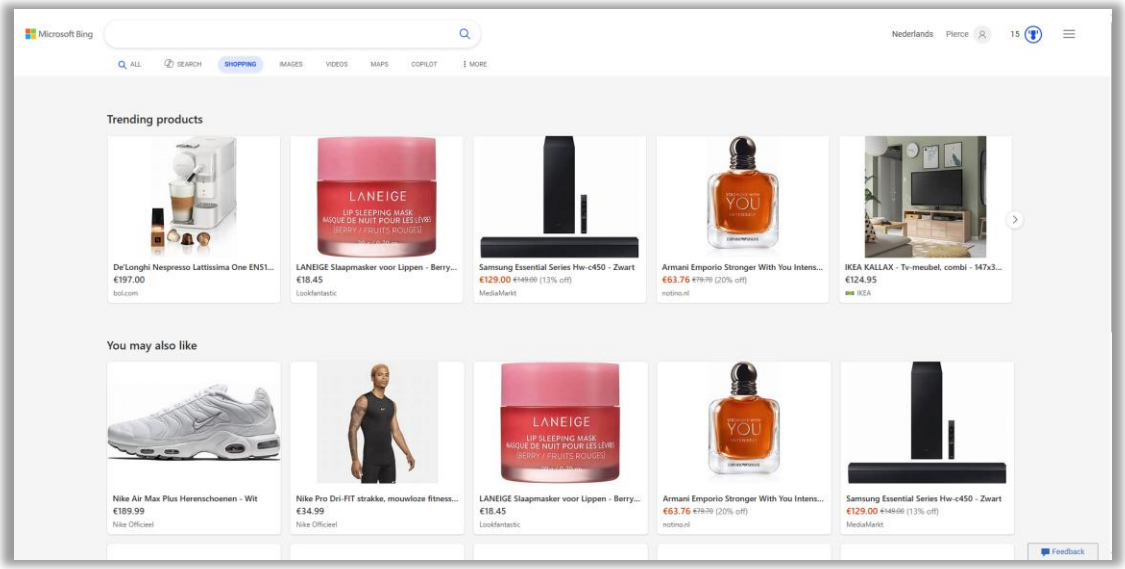
Figure 8: Bing Maps interface screenshot



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 GenAI 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.

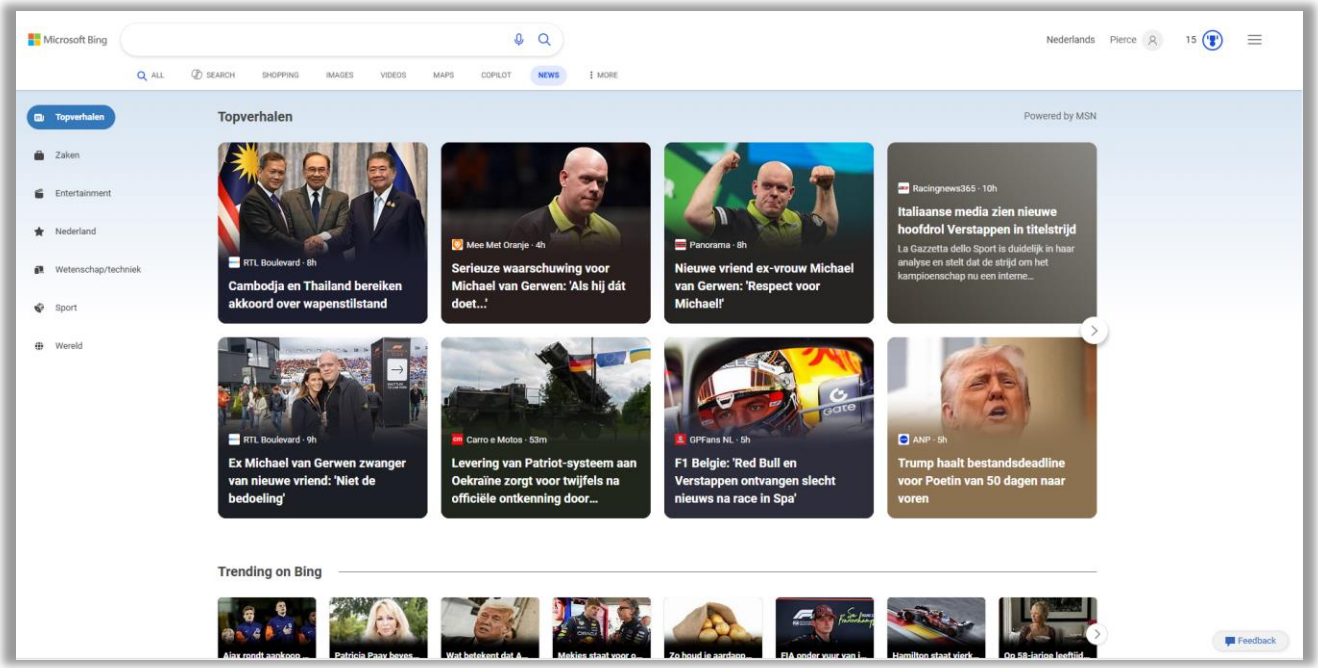
Figure 9: Bing Shopping landing page screenshot



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.

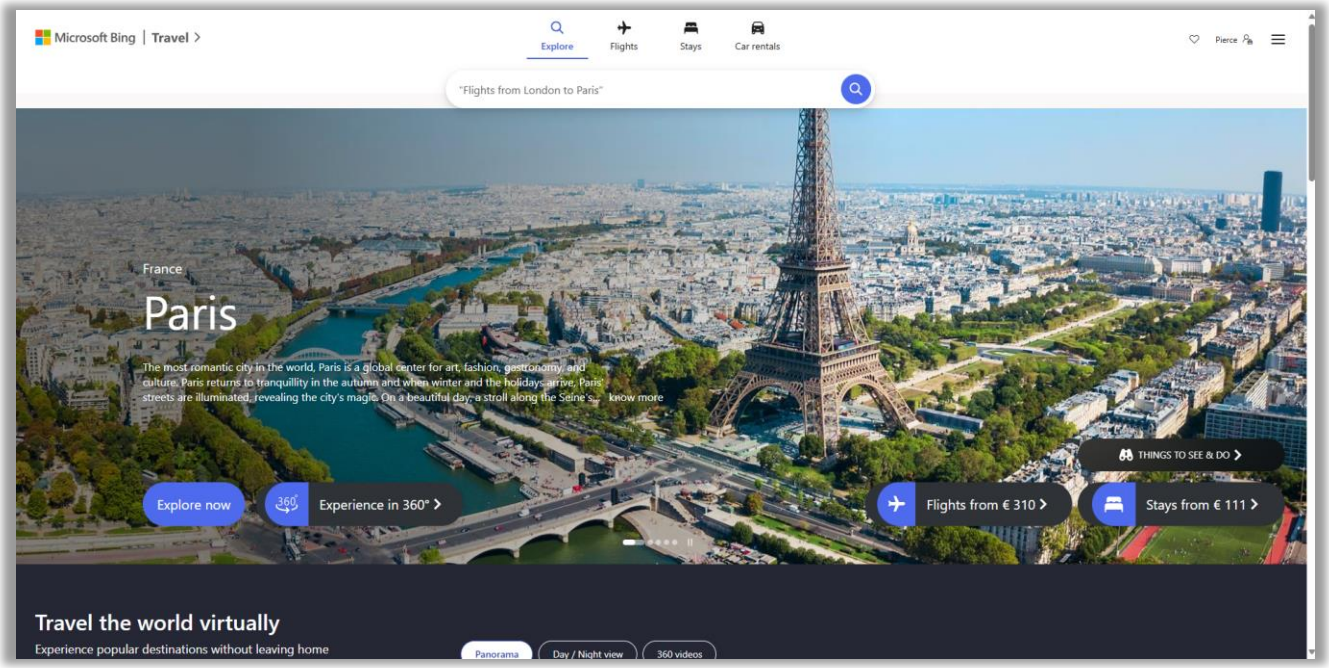
Figure 10: Bing News search vertical



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 11: Bing Travel home page screenshot



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.

Ancillary Search Features Risk Profile

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

Maps

- Risk that Maps could provide incorrect information related to polling locations.

- Risk that Maps show content that perpetuate bias, reinforce stereotypes, or discriminate against products, vendors, or businesses.
- Risk that Maps recommends sites and destinations where illegal, dangerous, or regulated goods or services can be acquired.
- Risk that Maps' aerial and street level imaging of sensitive locations, including military bases, may expose national security assets.
- Risk that images of children which do not conceal their privacy are shown on the “Travel the World Virtually” or in the Maps features.

Shopping

- Risk that Shopping over moderates and removes available products that are related to certain political parties or figures in a manner that perpetuates political bias.
- Risk that users could input targeted query seeking products in Shopping with slogans or graphics that include discriminatory content and view such products in the first few pages of results.
- Risk that Shopping shows recommended products that include graphics or logos that are violent, gory, vulgar, or profane.
- Risk that Shopping results link to websites related to the sale of illegal, dangerous, infringing, or regulated goods.
- Risk that Shopping listings show products with logos or graphics that promote/include gender-based violence.
- Risk that activities on Shopping are used to collect funds used to finance terrorist organizations.

News

- Risk that News returns low authority or low-quality news sources with election misinformation.
- Risk that third party web content linked in News results includes hate speech or discriminatory content or content that perpetuates hateful attitudes or harmful stereotypes or otherwise facilitates discrimination towards individuals or groups.
- Risk that in order to comply with government orders, News may block access to third party web content which could result in unnecessarily censoring legitimate access to information.
- Risk that News results include content that is sexual, pornographic, or promote legal prostitution.
- Risk that News shows content that instructs on, promotes, normalizes, or praises illegal materials.
- Risk that News results show content that content that instructs on, promotes, normalizes, praises, or facilitates promotes suicide, self-harm, or causing harm to others.
- Risk that News shows content containing inaccurate or misleading information about public health issues, medical conditions, treatments, side effects, public health orders or decrees, or other information that could potentially undermine public health.
- Risk that News results include misinformation related to crisis events or issues of general public security.

Travel

- Risk that Travel provides insufficient information about travel product/services sellers, preventing informed consumer decisions.
- Risk that Travel shows locations, listings, or destinations that are used for consensual, legal prostitution.

- 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 private property is featured as a destination or things to do in Travel results.
- Risk that Travel offerings, such as coupons, expose users to phishing, hacking, or data breaches.
- Risk that Travel shows locations or rental listings that are targeted for terrorist training, funding, or support for terrorist acts.

Advertisements on Bing

- Risk that political advertisements target vulnerable Bing users with election-related misinformation or exhibit a lack of transparency in political advertising targeting practices that can mislead voters.
- Risk that Advertisements include content that misleads consumers about the legitimacy of businesses, products, or services.
- Risk that Advertisements on Bing are biased in their targeting, affecting protected groups' ability to access critical services.
- Risk that sexual or pornographic ads appear on Bing.
- Risk that Advertisements promote illegal content or activities.
- Risk that Advertisements appear on Bing that promote/show addictive or harmful products or illegal substances that can negatively impact mental or physical wellbeing.
- Risk that Advertisements on Bing are used to promote legal but harmful substances.
- Risk that Advertisements on Bing expose minors to harmful activities targeted for child sexual exploitation and abuse.

The key mitigation measures Bing has implemented to address these potential risk manifestations are described in Bing's Approach to User Safety and in Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations.

Changes to 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, or adjusted as Bing seeks to improve users' experiences. Throughout the year, Bing tracks updates to features, including rollbacks, changes, and launches through its pre-launch risk assessments, RAI Review, and Shiproom processes.

In addition to these core business processes, Bing Compliance Oversight and Defensive Intelligence Response and Transparency ("DIRT") teams have implemented an additional process to identify product launches that may be in scope for a Critical Impact Risk Assessment ("CIRA"), which evaluates potential impact of the launch of new feature or feature update on Bing's DSA systemic risk profile.

Feature updates since the 2024 Systemic Risk Assessment report are discussed below.

Microsoft Copilot Consumer Chat and Copilot Search

As noted above, Bing is continually innovating and evolving the experiences to delight its users. The evolution of Bing's Generative AI ("GenAI") offerings since the 2024 DSA Systemic Risk Assessment represent the dynamism of search and Generative AI. Changes occurred in October 2024, when the consumer chat experience, known as Bing Chat, was separated from the Bing service. Bing Chat, as well as broader conversational AI stack, became a standalone experience called Microsoft Copilot Consumer Chat,

available at copilot.microsoft.com. Bing also launched Bing Generative Search in October 2024, which was rebranded as Copilot Search in April 2025.

Copilot Search, Bing's generative AI search experience, gives users quick, summarized answers with cited sources and suggestions for further exploration, making it easier than ever to discover more. Copilot Search enhances users' search engine results page using GenAI, creating a curated experience tailored to users' queries. Copilot Search is grounded with real-time insights, and defaults to authoritative search results for topics that are not supported.

Copilot Search also incorporates enhanced safeguards, including policies limiting GenAI responses on sensitive topics, real-time classifiers that flag and filter problematic or sensitive information, and embedded RAI instructions via meta prompting that refuse to answer harmful or illegal queries.

Copilot Search includes mechanisms to monitor risk and assess the effectiveness of its safeguards. User feedback is collected in-product and regularly reviewed by a designated team to identify and address emerging issues. Bing also maintains social listening channels to identify quality or safety issues in public discourse.

This integration of GenAI within the Bing suite of services is aligned with Bing's *Trustworthy Search Principles* to ensure the careful balancing of users' fundamental right to access information with important societal values for privacy, security, and safety.

Bing Video Creator

Bing Video Creator, launched in late May 2025, is a similar product to Bing Image Creator in that it generates videos without audio based on the prompt the user inputs. This feature builds on the success of Bing Image Creator and reflects Bing's continued investment in multimodal GenAI capabilities. Bing Video Creator is designed to enhance user engagement and creative expression by offering a visual storytelling format that can be used across education, content creation, and communication contexts.

While the feature is still in early stages, it introduces new content generation dynamics that are monitored closely through existing safety frameworks. This includes alignment with Bing's RAI policies, prompt moderation, and integration with content filtering systems to mitigate the risks of misuse, such as the generation of misleading or sensitive visual material. By expanding into video content, Bing is broadening the scope of its generative service while maintaining a focus on safe and responsible deployment.

Real Estate

Bing officially deprecated its global Real Estate vertical at the beginning of 2025, as part of a broader strategic realignment aimed at consolidating efforts around core search functionalities and emerging GenAI capabilities.

Systemic Risk Assessment Methodology

Methodology Overview

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 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 and Information
5. Human Dignity
6. Illegal Content and Activities
7. Mental and Physical Wellbeing
8. Private and Family Life
9. Protection of Personal Data
10. Public Health
11. Public Security
12. Rights and Protections of Minors

The Bing Risk Assessment team collected multi-disciplinary stakeholder input and information on updates to Bing features since July 2024, 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 internal 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 stakeholder validation of mitigations.
2. Guidance from authoritative sources, including the European Commission, on the severity of systemic risks.
3. Collection of Bing's published policies and official communications.
4. Bing internal stakeholder summarized policies and procedures.
5. Notes from internal consultations.
6. Notes from external consultations.
7. Bing internal metrics.
8. Transparency Reporting metrics.
9. Open-source data on public discourse related to Bing and systemic risk areas.
10. Bing products and features reviewed through the CIRA process
11. DSA risk and control matrix.

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 occurring on Bing through the potential 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 inputs from the definition and discovery phase to inform ratings and implemented a traditional risk assessment equation to aid in prioritization of risk for additional enhancements to mitigation measures.

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 for the DSA systemic harm areas, and content moderation metrics, to inform the relative probability of the risk.

The Bing Risk Assessment team highlighted DDR for the DSA systemic harm areas in the individual risk analysis sections below, as Bing teams track and monitor these metrics to measure overall product health and evaluate the effectiveness of Bing's multi-layered safety system.

- DDR, including DDR for the DSA systemic harm areas, 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 focuses on a very small portion of high-risk queries, over indexing on risky queries over non-risky queries in order to allow for more nuanced risk tracking.
- 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 search interventions***. 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 continues to focus on objectivity, using available data as part of a system-based assessment of severity. 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 ability to remediate) of impact to arrive at a severity score and rating.

Risks with irremediable impact on a greater number of systems on a broader scale will receive the highest severity ratings. 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 area 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

The Bing Risk Assessment team organized existing mitigations and controls according to industry-standard best practices and evaluated their implementation according to the Digital Trust & Safety Partnership ("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 Tooling

Risk Assessment Workbook

The Bing Systemic Risk Assessment Workbook is the primary analytical tool supporting Bing's 2025 Systemic Risk Assessment Report. It is designed to provide a structured, transparent, and comprehensive view of systemic risks across the Bing product ecosystem. The workbook integrates quantitative scoring methodologies with qualitative risk narratives to assess how specific risk scenarios may manifest across products, the likelihood and severity of those risks, and the effectiveness of existing mitigations.

Organized into results, methodology, assessment, and scoring components, the workbook includes detailed registers of risks, products, and mitigations, as well as scoring worksheets that calculate risk ratings using standardized scales.

Risk Assessment Results

Overall 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 ratings, and included the key mitigations relevant to that Risk Area. The rating scales and definitions are included in 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.

In the 2025 DSA Systemic Risk Assessment, the Residual Risk Ratings for Bing remain Low or Minimal across most assessed areas. However, the two risk areas, Illegal Content and Activities, and Human Dignity, are rated as Moderate. These areas continue to reflect an appropriate balance of fundamental rights and interests consistent with the unique risk profile of a search engine. While Bing assesses that current mitigations are effective, the service acknowledges the evolving nature of these risks and is committed to ongoing refinement of its mitigation strategies to maintain compliance with its systemic risk obligations under the DSA.

Figure 12: 2025 DSA Systemic Risk Assessment Ratings

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

Changes to Bing's Risk Profile

Over the past year, Bing has continued to operate steadily within a dynamic and evolving risk environment of technological, geopolitical, and social changes. Its risk profile remains stable even as new technologies emerge, and digital threats continue to evolve. As a VLOSE, its core function to enable users to access high quality information has not fundamentally changed, even with the integration of GenAI to make that more efficient.

This consistency within Bing's ecosystem, combined with the steady application of safety and relevance principles, contributes to a stable systemic risk profile. While not indicative of a change in the risk profile itself, Bing has expanded its risk capabilities by monitoring DDR aligned with DSA systemic risk areas. This enhancement enables more tailored analysis and reporting of harm-specific trends and mitigations over time.

One area of notable investment has been Bing's response to the growing concern around image-based abuse and the non-consensual sexual exploitation of individuals. Bing developed new NCII policies that apply across its ecosystem, as well as interventions to improve its abilities to detect and manage potential NCII, including classifiers to identify harmful prompts and potentially harmful content. Bing Image Creator downloads the StopNCII initiative's database of PhotoDNA hashes for reported and verified NCII content and continuously scans and removes NCII content from Bing's image index. Bing also updated the report a concern interface to include specific NCII reporting. Even with these investments, Bing recognizes the need and is continuing to explore ways to improve its ability to manage the complex risks associated with NCII.

Bing also implemented safeguards against prompts related to grooming and the generation of sexualized content involving children. These efforts are part of an evolving framework that incorporates human review, clear policy enforcement, and low-latency user appeals.

Bing has also made investments in the layered defenses for its Bing's GenAI capabilities, including bolstering upstream filters to detect jailbreak attempts, midstream blocking of known high-risk prompts, and downstream controls that restrict the generation of content related to public figures or harmful subjects. These targeted mitigations are complemented by continuous measurement pipelines. Bing uses both representative (realistic distribution of risky and non-risky queries) and adversarial (focused only on risky queries) datasets, as well as real-time social media monitoring, to detect emerging abuse patterns. These safeguards are regularly refined to ensure they remain effective without compromising user access to information.

As the service continues to evolve, Bing assesses how changes may affect systemic risks. The evolution of Copilot Search has been accompanied by careful evaluation of potential risk factors, with product design focused on respecting user rights to freedom of access to information as well as providing high authority content through grounding. Many core search functions remain unchanged, even with further enhancements to generative capabilities, and overall risk levels have not materially shifted. Bing continues to monitor external risk signals and internal safety metrics to ensure its mitigation strategy remains proportionate, resilient, and aligned with user rights and expectations.

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 2025

Systemic Risk Assessment, the Bing Risk Assessment team considered a variety of relatively higher risk events to determine whether each warranted a shift in the Inherent Risk rating.

These relatively higher risk events include continued geopolitical unrest and active armed conflicts and the continued expansion of the use of GenAI capabilities for both malicious and beneficial purposes.

The environment for VLOSEs has also changed over the past year. Most notably, there has been a rise in NCII, a decrease in public concerns in the EU around health misinformation, particularly compared to the heightened concern during the COVID-19 pandemic, and a relative decline around elections-based risks following the 2024 "year of elections". Risks related to violent extremism have remained relatively steady, though the groups and focal points continue to evolve. Bing has observed and is aware of these shifts and continues to monitor how they may impact its ecosystem. That said, Bing has not identified any significant impact on its own risk profile as a result of these broader global changes.

Over the past year, there have been multiple high visibility instances of AI-generated deepfakes often involving lewd portrayals or portrayals related to politics and elections. Bing is aware of the risks related to both AI generated deepfakes surfacing in Bing search results, as well as Bing users using Bing Image and Video Creator to create deepfakes.

Bing has implemented safeguards, such as C2PA enabled signing and the Content Credentials tool to help review and verify the authenticity of digital content. This system is part of a broader initiative to combat inauthentic content and ensure transparency in content creation and editing. Bing also implements the system of mitigations for Bing Gen AI experiences as discussed above. With respect to EU specific risks, Bing has not observed risks disproportionality affecting vulnerable or EU-based users.

Due to the dynamic and ever-evolving nature of online safety, there have been some shifts in the global risk

Changes in Risk Rating

Figure 13: Changes from 2024 to 2025 Systemic Risk Assessment Ratings

Risk Area	Residual Risk Rating
Freedom of Expression and Information	**Minimal -**
Human Dignity	Moderate
Illegal Content and Activities	**Moderate +**
Private and Family Life	**Moderate +**
Protection of Personal Data	Low
Civic Discourse and Electoral Processes	Low
Consumer Protection and Fraud	Low
Discrimination and Hate	Low
Mental and Physical Wellbeing	**Low -**
Public Health	Low
Public Security	**Low -**
Rights and Protection of Minors	**Low -**

Key: + indicates an increase of a rating from last year's assessment by one rating level, - indicates a decrease of a rating by one rating level

For Residual Risk Rating, there was a reduction from four Moderates in 2024 to two Moderates in 2025. From the 2024 to 2025 SRA, there are three reasons for changes in the risk ratings, (1) changes that are clerical in nature, (2) changes in the information environment, and (3) changes resulting from enhancements to Bing's mitigation measures.

With respect to clerical changes, the Illegal Content and Activities Residual Risk Rating increased from Low to Moderate over the past year primarily due to the internal reassignment of risks related to CSAM from the Rights and Protection of Minors risk area to the Illegal Content and Activities risk area. The Residual Risk Rating assigned to Rights and Protection of Minors also consequently decreased from Moderate to Low. This change better aligns the risk with legal obligations and enforcement priorities. Also, the Residual Risk Rating for Mental and Physical Wellbeing has decreased from Moderate to Low. This shift is based on internal and external indicators showing a lower likelihood of systemic harm. This is partly due to the passage of time since the launch of Gen AI outputs, which, while initially generating heightened scrutiny, have since become more normalized, and thus less likely. Bing recognizes that concerns remain and continues to invest in mitigation measures to safeguard and support user wellbeing considering an ever-changing risk area and product offerings.

Due to changes in the information environment in the EU and the world, the Residual Risk Rating for Public Security decreased from Moderate to Low based on a lower probability rating for Public Security in 2025. While still present within the EU information environment, public discourse related to the Israel-Gaza War has decreased in frequency since the 2024 assessment period, which impacted the Public Security probability rating. While risks related to terrorism typically require a longer time horizon than one year to materially decrease, they may swiftly and unexpectedly increase. Therefore, Bing is not changing its approach toward mitigating risks related to terrorism and continues to make investments in relevant mitigations.

Bing adjusted the Freedom of Expression and Information Residual Risk rating from Low to Minimal for this assessment. A significant factor in this shift is the removal of Chat from Bing and the creation of a separate Microsoft Copilot Consumer Chat product, as well as the Generative Search experience, recently rebranded as Copilot Search, that falls back to core search for complex or potentially harmful queries.

The rating changed because Copilot Search prioritizes providing users access to high quality information and, for queries that potentially present a higher risk of unexpected exposure to harmful content, defaults to providing authoritative search results rather than providing generative responses. This is in contrast with Microsoft Copilot Consumer Chat which, part of which was formerly Bing Chat, which blocks the generative responses for queries that present a higher risk of unexpected exposure to harmful content. Bing also continues to invest in monitoring overblocking for Bing Image and Video Creator results, to continually strive for more consistent and fair experiences.

The remaining risk ratings are consistent with the prior year assessment, which in some cases also reflects enhancement of mitigation measures to address risks that increased in probability during this assessment period. Bing invested in mitigation measures related to NCII across products, but industry developments and public concern related to NCII increased in the past year, therefore Bing assessed Private and Family Life to be a Moderate Residual Risk.

Results per Risk Area

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. In most cases, rather than removing or blocking content, Bing informs users of certain risks through public service announcements or warnings or provide users with options for tailoring their content.

Following Bing's principles, Bing's safety systems are balanced with the rights to Freedom of Expression and Information. Bing provides transparency around Bing's safety practices, content moderation processes and content removal data. For GenAI 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.

As previously mentioned within the *Changes in Risk Rating* section, 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 "Minimal"(See *Changes in Risk Rating* section for further details).

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Freedom of Expression and Information	Remote	High	Low	Optimized	Minimal

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 right to Freedom of Expression and Information, including the freedom and pluralism of the media occur on the service.
Risk Scenarios	• Over-moderation or overblocking of user-generated content online, over-removal, unjustified removal, or restriction of expressive content • Overly restricting access to information online, over-removal or restriction of access to lawful information

	• Recommendations that suppress pluralism • Inadequate internal safeguards against undue government or law enforcement influence over online content or safety interventions • Content or activities that chill public discourse, participation in online spaces, or open expression online • Content or activities that otherwise negatively impact freedom of expression, information, or pluralism
Theoretical Risk Manifestations absent Mitigations	• Risk that search over moderates the search results, removing or downranking URLs limiting access to information. • Risk that search's ranking processes disproportionately favor or demote certain types of content or viewpoints. • Risk that Bing's internal safety interventions could prevent individuals from accessing or generating information important to engaging in civic life, pursuing interests, or exercising other rights, or such safety interventions may be applied in biased ways. • Risk that Bing may unnecessarily reduce or block access to content based on government or law enforcement requests. • Risk that Shopping over moderates' available products (e.g., with slogans or graphics that may seem offensive). • Risk that News results might lead users to third party web content that is biased. • Risk that in order to comply with government orders, News may block access to third party web content which could result in unnecessarily censoring legitimate access to information. • Risk that Bing Image or Video Creator disproportionately block certain prompts or generated responses • Risk that the mitigations in Bing Image or Video Creator (such as blocklists, filtering, classifiers, metaprompts and other safeguards) disproportionately restrict certain types of content users seek • Risk that Bing Image or Video Creator disproportionately generates images that align with particular cultural standards or viewpoints, resulting in less diverse representations of the population • Risk that Bing Image or Video Creator limits artistic and journalistic expression by favoring certain prompts over others • Risk that Bing Image or Video Creator over moderation practices could lead to the prevention of image generation, even with lawful and fair prompts, thus restricting visual expression • Risk that GenAI Protection Systems flag and block the generation of legitimate content

Inherent Risk

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 "Remote."

Bing considered internal metrics and social listening data as factors in assessing a "Remote" level of probability relative to other potential systemic risks on Bing.

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 websites 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 GenAI features and that Copilot Search will return answers from core search if it cannot generate a GenAI response to the query.

The **Inherent Severity** of 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 risk area, 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 risk area is not considered significant.

With the rating for Inherent Probability as "Remote" and Inherent Severity as "High," the **Inherent Risk** associated with Freedom of Expression and Information on the Bing service is "Low."

Maturity of Mitigations

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 User Safety, in Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations summarized below.

The mitigations Bing has implemented follow best practices that promote trust and safety in every aspect. 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 down to "Minimal."

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 GenAI features that offer new ways for users to approach research, learning, and search. The highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to Freedom of Expression and Information and expansions on previously described cross-platform mitigations with specific applications to Freedom of Expression and Information. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated freedom of access to information. This layered approach ensures both platform-wide consistency and harm-specific responsiveness. For ***prioritizing high authority content in search,*** 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.

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's ***Transparency Reports*** as applied to Freedom of Expression and information, further provide transparency on content moderation.

As an additional quantitative signal of safety performance, Bing ***monitors overblocking metrics*** (identifying where Bing Image or Video Creator has unnecessarily blocked a creative response) in order to continually to assess the effectiveness and appropriateness of its GenAI safety measures to appropriately prioritize users' fundamental rights to information.

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.

Microsoft also is a member of the ***Global Network Initiative ("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.

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 unexpected exposure to content that could negatively affect users' 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. GenAI 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," consistent with the rating assigned in 2024.

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Human Dignity	Likely	Critical	High	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 Scenarios covered by this Risk Area	• Human trafficking (other than sex trafficking) or coercive conduct depriving others of their human rights for personal gain • Sexual exploitation, non-consensual prostitution or pornography, or sex trafficking, and extreme pornography • Graphic violence, human gore, vulgarity, profanity, or other content that demeans, shocks, or offends • Sexual or pornographic content, and consensual, legal prostitution • Content or activities that otherwise negatively impact Human Dignity

Theoretical Risk Manifestations absent Mitigations	• Risk that search engines return content that facilitates or depicts human trafficking in the top search results. • Risk that search engines return content that instructs on, promotes, normalizes, or praises human trafficking other than sex trafficking in the top search results. • Risk that search engines return content that can be used to access the services of individuals being trafficked in the top search results. • Risk that search results or autosuggest features lead users to web results that aid human traffickers by promoting false job postings. • Risk that search engines return content that is graphic, violent, gory, vulgar, profane or that otherwise demeans, shocks, or offends in the top search results. • Risk that search engines return content that is sexual or pornographic or link to information on locating consensual and legal prostitution in the top search results • Risk that Maps or Travel show locations, listings or destinations that are used by malicious actors for human trafficking • Risk that Maps or Travel show locations, listings or destinations that are used by malicious actors for sexual exploitation or non-consensual prostitution • Risk that News results include content that is sexually exploitative, graphic, pornographic, violent, vulgar, or profane. • Risk that Shopping shows recommends products that include graphics or logos that are violent, gory, vulgar, or profane. • Risk that News results include content that is sexual, pornographic or promote legal prostitution. • Risk that sexual or pornographic ads appear on Bing. • Risk that Maps and Travel show locations, listings or destinations that are used for consensual, legal prostitution. • Risk that Bing Image or Video Creator is used to create materials that depict sexual exploitation, pornography, or extreme pornography • Risk that Bing Image or Video Creator is used to create violent, gory, or profane content • Risk that Bing Image or Video Creator is used to generate images that are used to recruit or exploit potential trafficking victims • Risk that Bing Image or Video Creator is used to generate or manipulate images that are sexually exploitative, contain pornographic content, or promote prostitution • Risk that Bing Image or Video Creator facilitates the generation of explicit adult content involving real or synthetic individuals without consent • Risk that Bing Image or Video Creator is used to generate manipulate images that are violent, gory, vulgar, or include profanity • Risk that GenAI summaries are based on low authoritative sources and produce content that may be harmful or misleading. • Risk that adversarial user queries or grounding data may bypass the safety features for Copilot Search and produce harmful or misleading content

Inherent Risk

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 "Likely."

Bing considered internal metrics and social listening data as key factors in assessing a “Likely” level of probability relative to other potential systemic risks on Bing.

The **Inherent Severity** of content and activities that negatively impact Human Dignity is rated as “Critical” primarily due to the gravity of risks within this risk area, considering the potential for significant harm to wellbeing, societal, economic, and security systems at the individual and potentially broader level. Some inherent risks within this risk area, such as human trafficking and prostitution, have potentially irremediable

impacts. Other risks, such as gore, vulgarity, profanity, or pornographic content, may be remediable but can be significant.

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

Maturity of Mitigations

Bing has implemented a robust set of mitigations to address risks related to Human Dignity, including the mitigations that apply across harms as described in Bing's Approach to User Safety, and inventoried in Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations, and well as risk-specific mitigations 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 down to "Moderate," as detailed above.

While not a proxy for Residual Risk, Bing has measured the DDR for Human Dignity on Bing Search as an average of 1.28% from September 2024 to April 2025. This means that Bing estimates 1.28% of search traffic may include content that may have negative effects related to Human Dignity.

Bing nevertheless, continues to invest in developing and enhancing strategies to further mitigate and manage risks related to Human Dignity. The highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to Human Dignity and expansions on previously described cross-platform mitigations with specific applications to Human Dignity. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with degrading, explicit, or exploitative content. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

To uphold the principle of Human Dignity, Bing has implemented ***safety measures for search suggestions*** aimed at curbing the generation of suggestions that might tarnish this fundamental right. This also includes creating avenues for users to effortlessly report suggestions they deem inappropriate, ensuring such concerns are promptly addressed.

As part of the ***Safety Measures – Defensive Search Interventions*** mitigation, Bing has a dedicated team accountable for implementing algorithmic defensive search interventions (as well as metrics monitoring and remediation) to address high impact issues in search results, such as information manipulation, hate speech, and other problematic content that could negatively impact user safety.

Bing uses query classifiers to identify queries that may be associated with elevated risk, such as those that could surface degrading, harmful, or exploitative content, and those trigger a set of defensive search interventions. These algorithmic measures are integrated rather than standalone and work together to uphold Bing's relevance, quality, and credibility standards—particularly in cases where harmful content may exploit data voids or manipulate user intent.

When a query is flagged as potentially risky, Bing may respond with measures such as boosting high-authority sources or applying algorithmic dampening to deprioritize low-quality or harmful content. These interventions are applied in accordance with Bing's core search principles and are actively monitored by the dedicated team that manages implementation and assesses impact across high-risk areas. This combined approach allows Bing to detect potentially harmful search patterns early and respond with targeted mitigations that preserve both user safety and access to high-quality information.

For ***safety features on visual search***, Bing also uses the hash-matching technologies PhotoDNA and MD5 to detect matches of previously identified CSEAI in these images provided by users to prevent upload. In the context of the search, the use of these technologies furthers Bing's goal to avoid 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 image upload, Bing helps to ensure that CSEAI is not included in training data.

Microsoft regularly engages in ***social listening for GenAI features*** to review the prevalence of issues within Bing Image or Video Creator 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 potentially impacting Human Dignity.

Users can use "***report a concern***" (See *Content Moderation* section for further details) to report unexpected offensive or harmful material (e.g., reporting unexpected adult, violent, or gore in results, non-consensual intimate imagery, or Harm - CSEAI).

The ***Terms and Conditions for Users***, found within the Microsoft Service Agreement (MSA), as applied to Human Dignity prohibit using Microsoft services to publicly display, generate, or share inappropriate content or material (involving, for example, nudity, bestiality, pornography, offensive language, or graphic violence). The Bing Image Creator Terms of Use further prohibit use of Bing's GenAI features to create or share adult content, violence or gore, or content that is otherwise disturbing or offensive.to create or share adult content, violence or gore, or content that is otherwise disturbing or offensive.

Illegal Content and Activities

Risks related to Illegal Content and Activities include CSEAI, potential IP infringement, defamation, and the promotion or sale of illegal, dangerous, or regulated goods or services.

Bing takes proactive steps to prevent the dissemination of CSEAI using PhotoDNA scanning. Bing's content moderation systems are designed to process requests regarding illegal content, with reviews conducted by the appropriate teams. While Bing employs automated content detection (ACD) to identify and exclude from the index certain high-risk content, most notably child sexual exploitation and abuse imagery (CSEAI) and, more recently, NCII, the majority of content removal decisions rely on a report-and-react model.

For most types of harmful or policy-violating content, automated detection is not feasible due to the high degree of contextual nuance involved. Instead, Bing responds to reports through established content moderation systems and service-level agreements (SLAs), in accordance with its notice-and-takedown framework. Bing includes information on illegal content removal in transparency reports.

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 "Moderate" (Refer to *Changes in Risk Rating* above for further details).

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

Risk Definition

Risk Area	Illegal Content and Activities
Risk Definition	Risk related to the conduct of illegal activity or dissemination of illegal content through the service.
Risk Scenarios	• IP infringement, defamation • Promotion, concealment, acquisition, or sale of illegal, dangerous, or regulated goods (including counterfeit goods, animals, drugs, psychoactive substances, criminal property, firearms, and other weapons) and services relevant to the same • Facilitation of unlawful immigration • Causing suffering of animals • Creation, dissemination, or promotion of CSEAI • Content or activities that otherwise disseminating illegal content or enabling illegal activity through the service
Theoretical Risk Manifestations absent Mitigations	• Risk that search engines return content that infringes on IP rights or disseminates defamatory statements in the top search results. • Risk that search engines return content that contains illegal content sources, such as third-party websites offering illegal content streaming or music downloads or sites providing counterfeit goods or other deceptive commercial items for sale in the top search results. • Risk that search engines return otherwise illegal content in the top search results. • Risk that search engines return content that includes CSAM or CSEAI or applications, tools, or services for creating synthetic CSEAI. • Risk that Shopping results show counterfeit or knock-off products that infringe on the IP of brands. • Risk that Shopping results link to websites related to the sale of illegal, dangerous, infringing, or regulated goods (including illegal drugs, weapons, psychoactive substances, pharmaceuticals, etc.) • Risk that Advertisements promote illegal content or activities. • Risk that Maps recommends sites and destinations where illegal, dangerous, or regulated goods or services can be acquired. • Risk that News shows content that instructs on, promotes, normalizes, or praises illegal materials. • Risk that Copilot Search or Bing Image or Video Creator are used to generate third party IP without authorization in generated responses • Risk that Bing Image or Video Creator are used to generate third party IP without authorization in generated responses • Risk that Bing Image or Video Creator is used to generate information or images that encourage animal cruelty • Risk that Bing Image or Video Creator is used to create images that promote or glorify illegal activities, products, or substances

Inherent Risk

The **Inherent Probability** of systemic risks related to Illegal Content and Activities appearing on deriving from use of Bing's products and services absent sufficient mitigations is assessed as "Likely".

Bing considered internal metrics and social listening data as key factors in assessing a "Likely" level of probability relative to other potential systemic risks on Bing.

The **Inherent Severity of** 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 risk area 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 risk area 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."

Maturity of Mitigations

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 User Safety, in Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations 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 down to "Moderate."

While not a proxy for Residual Risk, Bing has measured the DDR for Illegal Content and Activities on Bing Search as an average of .14% from September 2024 to April 2025. This means that Bing estimates .14% of search traffic may include 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 highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to Illegal Content and Activities and expansions on previously described cross-platform mitigations with specific applications to Illegal Content and Activities. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with illegal content. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

In response to user search queries for CSAM, Bing displays ***in-product indicators*** and PSAs stating that CSAM is illegal; Bing 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 Search displays unrelated images from popular web content. For Bing Image and Video Creator, the response to CSAM related queries is disengagement stating that the request is inappropriate or harmful.

Similarly, Bing also states in its 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 account suspension***.

For ***incident response and 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, 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.

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.

The ***Report a Concern*** forms include functionality to report unlawful content.

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.

The ***Terms and Conditions for Users***, found within the Microsoft Service Agreement (MSA), as applied to Illegal Content and Activities prohibit using Microsoft services to do anything illegal, or try to generate or share content that is illegal; to publicly display, generate, or share inappropriate content or material (involving, for example, criminal activity); and to violate or infringe upon the rights of others (e.g., unauthorized sharing of copyrighted music or other copyrighted material). The Bing Image Creator Terms of Use further prohibit use of Bing's GenAI features to infringe on others' legal rights, including intellectual property rights; to create or share child sexual exploitation or abuse material; or to do anything illegal.

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.

Bing adheres to Microsoft privacy standards and relevant laws. Bing completes Data Protection Impact Assessments ("DPIA") 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.

Microsoft's longstanding belief that privacy is a fundamental human right has also informed every stage of Bing's development and deployment of the Copilot Search 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 Search. As Bing evolves its approach to providing the Copilot Search GenAI 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. 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," and furthermore, did not see a change in the scoring for Protection of Personal Data for since 2024.

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.
Risk Scenarios	• Collection, processing, access to, or release of information without consent or other misuse of data, including data mining or targeted advertisements without consent • Phishing, hacking, degradation or other abuse of service or data breaches or other unauthorized access, alteration, destruction, or disclosure of data or otherwise insufficient protection of data • Disclosure or exposure of personal information or data online • Other content or activities negatively impacting the Protection of Personal Data
Theoretical Risk Manifestations absent Mitigations	• Risk that users are harmed by inappropriate processing of personal data, such as lack of transparency, control, access to remedies, or disclosures. • 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 search history data, if inadvertently disclosed, reveals an individual's beliefs, relationships, sexuality, medical issues, or other private information. • Risk that search engines return content that enables 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, including credit card numbers, medical records, etc. in the top search results. • Risk that users of Ancillary Features may be harmed by inappropriate processing of personal data they've provided to the service. • Risk that Travel offerings, such as coupons, expose users to phishing, hacking, or data breaches. • Risk that personal information shared with Ancillary Features could be inadvertently disclosed or exposed online. • Risk that Bing Image or Video Creator is used to generate images that are used for phishing campaigns

Inherent Risk

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."

Bing considered internal metrics and social listening data as key factors in assessing a "Highly Likely" level of probability relative to other potential systemic risks on Bing. Public discourse reflects a higher level of concern related to the security of data provided to GenAI tools, which leads to the assessment of inherent probability as "Highly Likely." The **Inherent Severity** of content and activities that negatively impact the Protection of Personal Data is rated as "High" primarily due to the gravity of risks within this risk area, considering the potential for significant harm to economic, security, and societal systems at the individual and local level. Some risks within this risk area – 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."

Maturity of Mitigations

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 User Safety, in Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations 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 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 .02% from September 2024 to April 2025. This means that Bing estimates .02% of search traffic may include content that may have negative effects related to 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 highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to the Protection of Personal Data and expansions on previously described cross-platform mitigations with specific applications to Protection of Personal Data. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with unauthorized disclosure, phishing, or hacking. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

Bing 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. Unauthenticated users can control their search history through Bing's in-product controls using Bing's settings page.

Copilot Search has data retention and deletion policies to help ensure that personal data collected through this feature is kept as long as needed.

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. 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.

The Visual Search feature deploys a ***blurring 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. Access to most Bing Search history data is limited to Bing employees. 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.

For any new products or features that pose materially different personal data processing challenges, Bing completes or updates ***DPIAs*** as part of its rigorous privacy protocol. These assessments are reviewed by

Microsoft's Data Protection Officer and are subject to reevaluation 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.

Bing's ***Report a Concern*** tool includes options for users to report exposed personal or private information.

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.

Civic Discourse and Electoral Processes

Risks related to Civic Discourse and Electoral Processes 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 Assessment Period, there were elections in some European Union member states. 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.

This complexity is furthered by the continued rise in GenAI 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. Bing's core

ranking algorithms and systems include 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 pointing to authoritative local sources ahead of specific elections, rapid incident response processes, coordination (through Microsoft) with election authorities, and investment in GenAI 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," and furthermore did not see a change in the scoring for Civic Discourse and Electoral Processes for this year.

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.
Risk Scenarios	• Coordinated manipulation, manipulation, disinformation, inauthentic activity, misleading communications, digital threats, and foreign interference, including deceptive GenAI election content • Election misinformation • Political polarization, political bias, and the creation of echo chambers • Intimidation, harassment, coordinated targeting, or coercion of political or civic actors • Content or activities otherwise negatively impacting Civic Discourse and/or Electoral Processes
Theoretical Risk Manifestations absent Mitigations	• Risk that bad actors manipulate search results or exploit data voids to surface low authority, low quality content with an intent to deceive or manipulate audiences. • Risk that search engines return content that is 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 in the top search results. • 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 Bing's safety systems introduce bias or limit access to important election information if applied inconsistently. • Risk that search results or autosuggest features might lead users to web results that favor certain political views, influencing perceptions and polarizing suggestions related to elections. • Risk that Bing's partnerships with external experts or fact checkers introduce bias into the Bing service. • Risk that News results include disinformation, manipulated content, inauthentic content, content produced by malign foreign actors related to elections, electoral processes, voting procedures, election outcomes in EU member states and around the world; or content providing efforts to undermine election integrity. • Risk that News returns low authority or low-quality news sources with election misinformation.

	• Risk that political advertisements target vulnerable Bing users with election-related misinformation or exhibit a lack of transparency in political advertising targeting practices that can mislead voters. • Risk that Maps could provide incorrect information related to polling locations. • Risk that News returns news sources that increase political polarization and filter bubbles. • Risk that Shopping over moderates and removes available products that are related to certain political parties or figures in a manner that perpetuates political bias. • Risk that News inadvertently returns news articles that are targeted to damage a political actor's reputation or credibility. • Risk that bad actors misuse Bing Image and Video Creator to generate deceptive election content, such as deceptive images of political candidates, elections, and voting processes • Risk Bing Image and Video Creator exacerbates societal divisions through the creation of images that might exaggerate political stances, deepening echo chambers • Risk that bad actors misuse Bing Image and Video Creator is to generate images that are used to intimidate, harass, or coerce political or civic actors into complying with their demands • Risk that GenAI summaries are based on low authoritative sources and produce content that may be harmful or misleading • Risk that adversarial user queries or grounding data may bypass the safety features for Copilot Search and produce harmful or misleading content

Inherent Risk

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."

Bing considered internal metrics and social listening data as key factors in assessing a “Likely” level of probability relative to other potential systemic risks on Bing.

The Bing Risk Assessment team also considered the number of elections in Europe over the assessment period as a potential impact to the probability level. The Inherent Probability rating is lower this year due to the lower number of elections in Europe compared to last year.

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 risk area, 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 risk area 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.”

Maturity of Mitigations

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 User Safety, in Appendix II: Bing’s Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations summarized below.

The mitigations Bing has implemented relative to Civic Discourse and Electoral Processes follow best practices that promote 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 down to "Low."

As part of preparing for elections in the EU, Bing reviewed and implemented 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). Bing's existing practices reflected the guidelines, building on the additional improvements made in 2024, the "year of elections" in the EU. Microsoft and Bing have also undertaken multiple engagements with the Commission, press agencies, and Digital Service Coordinators.

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 .03% from September 2024 to April 2025. This means that Bing estimates .03% of search traffic may include 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 highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with elections, content integrity, and information manipulation. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

Bing takes a multi-layered approach to protecting election integrity, with mitigations across the various layers of the Bing Safety Funnel. Bing also 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.

Product Quality

The ***Terms and Conditions for Users*****,** found within the Bing Image Creator Terms, as applied to Civic Discourse and Electoral Processes, prohibit use of the service to engage in activity that is false or misleading; to attempt to create or share content that could mislead or deceive others, including for example creation of disinformation, or deceptive impersonation; to remove or alter content credentials or other provenance methods, marks, or signals that indicate that the Creations were generated by the service; or to otherwise attempt to mislead others about whether the Creations were generated by the service.

Furthermore, Microsoft ***does not permit political advertising*** within the Microsoft Advertising ecosystem, which includes Bing. Advertising policies prohibit ads for political candidates, parties, ballot measures, fundraising, or other election-related content globally. Ads that were issue-based and could be perceived as politically sensitive are also disallowed. Microsoft Advertising is a COPD signatory, and its specific commitments are outlined in Microsoft's most recent COPD report.

Microsoft's Content Integrity Tools, originally designed as a pilot program for the 2024 election cycle and to gain feedback about Content Credentials-enabled tools, allow users to add content credentials to their own authentic content, in effect ***Content Credentials as a Service*****.** The Tools were available to political

campaigns in the EU, as well as to elections authorities and select news media organizations in the EU and globally.

These tools included a partnership and collaboration with fellow Tech Accord signatory, TruePic. Announced in April 2024, this collaboration leveraged TruePic's mobile camera SDK enabling campaign, election, and media participants to capture authentic images, videos and audio directly from a vetted and secure device. Called the "Content Integrity Capture App" (an app that makes it easy to directly capture images with C2PA enabled signing) launched for both Android and Apple and can be used by participants in the Content Integrity Tools pilot program.

Product Safety

Bing has ***special informational and answer segments for*** **elections** including 'How to Vote' answers in defined markets to provide users with high-authority information related to the election, localized by language and member state. Bing Search directs users to official election sources, such as https://elections.europa.eu/, and other high-authority sites locally.

Beyond standard measurement to evaluate the quality of search results against our policies, and ahead of specific elections, Bing ***monitors elections-related DDR metrics*** through election-related query sets to measure the presence of content that goes against its search policies across Bing Search features including Web, Image, and Video. Bing may invest additional resources into evaluating the quality of search results through sample-based metric evaluations and 'red team' testing on Core and Copilot Search and/or creative content generation features.

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 took place both in product review settings and in broader cross-functional teams dedicated to elections at Microsoft.

Targeted Interventions

Microsoft identifies, tracks, and reports potential information operations and foreign malign influence campaigns, working with independent third-party partners and through the ***threat intelligence team focused on nation state actors and information operations***. Public reporting from that team on cyber influence operations can be found on their website at Microsoft's Threat Intelligence Team (MTAC).

Copilot Search and Bing Image and Video Creator utilize defensive interventions as well as local election mitigations, such as blocklists, to restrict generation of images or certain types of content concerning politicians in upcoming elections and to prevent responses to certain election-related queries out of an abundance of caution. Through Microsoft's dedicated cross-functional team focused on democracy and elections, these defensive interventions and election-related mitigations include information on political parties, candidates, and elections from local election authorities (including in the EU) and other high-authority third-party sources. Bing continues to invest in these mitigations, including through focused testing in key EU languages used by Bing users in the EU.

Bing uses ***social listening*** to evaluate online conversations around potential RAI risks related to election and political topics, enabling mitigation actions as necessary.

Content Moderation

Microsoft maintains an "***Election Communications Hub***" to support democratic governments and political parties worldwide as they build secure and resilient election processes. This hub gives election authorities the ability to rapidly report any issues or concerns to Microsoft security and support teams in the days and weeks leading up to their elections, allowing swift assistance with major security challenges.

Microsoft maintains a web page for ***deceptive AI election reporting***, where political candidates can report concerns about deepfakes on Microsoft services. This tool empowers political candidates globally to aid in detecting harmful deepfakes. The webpage is available and localized throughout the EU.

Research and Partnership

Microsoft undertakes significant efforts to support election integrity through a whole-of-company approach. Several cross-functional teams, including dedicated internal product reviewers, global and regional election working groups, and specialists in elections and democracy focus on election-related risks, policies, and AI-related issues. These teams regularly review product features and inform decisions across Bing Search, GenAI features, and other relevant services.

Microsoft has a global election working group that convenes stakeholders from Bing, local support teams, and broader Microsoft policy functions to share insights, identify risks, and coordinate mitigations related to civic integrity worldwide. These cross-functional ***teams at Microsoft dedicated expressly to addressing election-related concerns*** ensure Microsoft undertakes efforts to improve its processes to support election integrity using a whole-of-company approach.

A key component of the dedicated cross-functional team's strategy is establishing lines of communication between election authorities in EU member states and ***identifying risks to electoral processes,*** including potential foreign information operations targeting elections.

Microsoft maintains a "***Campaign Success Team***" to help political campaigns navigate cybersecurity challenges and the evolving landscape of AI. This team advises and supports campaigns in combating cyber-influence operations and protecting the authenticity of their content and images.

Microsoft implements ***political group awareness campaigns and trainings***, such as Deceptive AI and Elections Trainings for political parties and campaigns across the EU. These trainings help participants learn to identify deceptive uses of AI and misinformation in elections and report them using Microsoft's Deceptive AI Reporting Tool (discussed below).

Despite evolving threats to democracy, Microsoft continues to protect open and secure democratic processes by providing services and technologies to secure critical institutions, ***safeguard electoral processes from cyberattacks***, and build public trust in voting procedures.

Microsoft works closely with election-related customers to provide ***security guidance*** and tools to improve cyber-resilience and protect electoral integrity.

Microsoft supports the International Foundation for Electoral Systems ("IFES") in strengthening the cybersecurity practices of investigative journalists reporting on abuses of state resources in elections. Microsoft also partnered with the National Democratic Institute ("NDI") to strengthen cybersecurity infrastructure for political parties and campaigns internationally. Microsoft engages consistently with ***civil society and industry partners*** through a dedicated democracy and elections team. These partnerships included collaboration with International IDEA and the NATO Hybrid Centre of Excellence to identify and promote best practices for supporting election integrity.

For ***external engagements with governments***, 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 EU.

Additionally, Bing collaborates with Microsoft Research and external research organizations to support studies on safe product design, responsible AI, and information manipulation. In preparation for global elections, Microsoft Research conducts internal research on information integrity and elections in the era of GenAI.

Transparency and Reporting

Bing is a signatory to the ***EU COPD*** and as a signatory to the EU COPD, and as an active participant in the related working groups, Bing fulfills its commitments to reduce disinformation risks. This includes participation in the Elections Working Group's rapid response program, delivering transparency reporting, offering the Content Credential tool to help users assess the trustworthiness of websites or domains, and empowering them to make informed decisions about online content. Bing also provides enhanced reporting on elections designated as a "crisis" under the COPD framework.

As part of its commitments under the EU Code of Practice on Disinformation (COPD), Bing provides ***transparency reporting*** via biannual reporting under the Code on the systems and policies in place to prevent the spread of misinformation, including related to elections, across the EU. The biannual COPD reports include comprehensive information on measures related to declared crises and agreed-upon EU elections. Bing also serves as a joint coordinator in the Generative AI (GenAI) working group under the COPD.

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, and 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.

Bing recognizes that risks associated with fraud in advertising have been a major focus for regulators and platform providers like Bing in the last year, with concerns about how GenAI exacerbates these risks top of mind.

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" (See below for further details).

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Consumer Protection and Fraud	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.
Risk Scenarios	• Scamming, defrauding, spamming, use of malware, virus, or spyware, fraudulent business or advertising, or identity theft • False representation, abuse of position, falsifying information, including the creation of fake accounts, use of bots, or other deceptive commercial practices, including false claims and misleading advertisements

	• Insufficient information about sellers or products, preventing informed consumer decisions • Content or activities that otherwise negatively impact consumer protection
Theoretical Risk Manifestations absent Mitigations	• Risk that search results or autosuggest features lead users to web results that inadvertently prioritize scam businesses or fraudulent websites. • Risk that search engines return content that is misleading to consumers about the legitimacy of businesses, products, or services in the top search results. • Risk that users encounter third-party content including instructions on how to defraud users in the top search results. • Risk that Advertisements link to malware. • Risk that Advertisement include content that misleads consumers about the legitimacy of businesses, products, or services. • Risk that Shopping and Travel listings include fraudulent, deceptive, or unsafe products. • Risk that Maps or News results include content that misleads consumers about the legitimacy of businesses, products, or services. • Risk that Travel provides insufficient information about travel product/services sellers, preventing informed consumer decisions. • Risk that Bing Image and Video Creator is misused to generate highly realistic images of non-existent individuals or impersonate companies for use in phishing schemes, fake endorsements, or efforts to create synthetic identities that could be used to defraud businesses and consumers • Risk that Bing Image and Video Creator is misused to insert virus or malicious code into generated images to create malware that can be used for malicious activities • Risk that GenAI summaries are based on low authoritative sources and produce content that may be harmful or misleading. • Risk that adversarial user queries or grounding data may bypass the safety features for Copilot Search and produce harmful or misleading content

Inherent Risk

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 "Likely."

Bing considered internal metrics and social listening data as key factors in assessing a “Likely” level of probability relative to other potential systemic risks on Bing.

The **Inherent Severity** for fraudulent content or activities on the Bing service is rated as "High" due to the gravity of risks within this risk area 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 risk area 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 “Likely” and Inherent Severity as “High,” the **Inherent Risk** associated with Consumer Protection and Fraud on the Bing service is “High.”

Maturity of Mitigations

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 User Safety, in Appendix II: Bing’s Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations 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 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 .02% from September 2024 to April 2025. This means that Bing estimates .02% of search traffic may include 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 highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to Consumer Protection and Fraud and expansions on previously described cross-platform mitigations with specific applications to Consumer Protection and Fraud. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with scams, fraud, and false representation. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

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.

Bing's Webmaster Guidelines include details on prohibited practices, specific to spam and fraud, intended to manipulate or deceive the Bing Search algorithms.

Similarly, Bing also states in its 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 ***account suspension***.

Products in Shopping, primarily leverage product advertisements from Microsoft Advertising. For ***shopping content moderation***, 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.

For consumer protection with ***Microsoft Advertising Policies***, 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.

The ***Microsoft Advertising fraud 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 changes 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.

The ***Terms and Conditions for Users***, found within the Microsoft Service Agreement (MSA), as applied to Consumer Protection and Fraud prohibit using Microsoft services to send spam or engage in phishing, or try to generate or distribute malware or engage in activity that is fraudulent, false or misleading. The Bing Image Creator Terms of Use further prohibit use of Bing's GenAI features to engage in activity that is fraudulent, false, or misleading; to attempt to create or share content that could mislead or deceive others, including for example creation of content enabling fraud; or to remove or alter content credentials or other provenance methods, marks, or signals that indicate that the Creations were generated by the service.

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. GenAI features deploy additional mitigations to prevent harmful AI-generated outputs. Additionally, GenAI features undergo review prior to launch to identify the potential for harm, including harm related to Discrimination and Hate, under Microsoft's RAI program. 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	Not Likely	High	Moderate	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, impacting individuals and groups based on race, ethnicity, religion, gender, sexual orientation, or other protected traits, or reflecting, reinforcing, or perpetuating harmful stereotypes, biases, or inequalities.
Risk Scenarios	• Bias and discriminatory practices, features, safety interventions, or policies unfairly disadvantaging protected individuals or groups • Bias and discriminatory practices, user actions, activities, or behaviors negatively impacting protected individuals or groups • Hate speech and discriminatory content, content that insults, degrades, or dehumanizes protected individuals or groups, including hateful imagery • Content or activities that otherwise negatively impact the fundamental right to non-discrimination
Theoretical Risk Manifestations absent Mitigations	• Risk that Bing algorithmic systems contain bias that negatively impacts protected user groups on Bing. • 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 Bing's content moderation practices are implemented unfairly resulting in biased or discriminatory search results. • Risk that search results autosuggest features lead users to web results that could expose users, including minors, to hateful ideologies, beliefs, or symbols. • Risk that search engines return content that contain hate speech or discriminatory content in the top search results. • Risk that search engines return content that includes recorded livestreams from hate-fueled violent attacks in the top search results. • Risk that search engines return content that provides hateful rhetoric or advocates the use of violence against people because of their race, or other protected traits in the top search results. • Risk that advertisements on Bing are biased in their targeting, affecting protected groups' ability to access critical services. • Risk that Shopping, Maps, and Travel show content that perpetuate bias, reinforce stereotypes, or discriminate against products, vendors, or businesses. • Risk that Shopping shows products with slogans or graphics that include hate speech or discriminatory content. • Risk that users could input targeted query seeking products in Shopping with slogans or graphics that include discriminatory content and view such products in the first few pages of results. • Risk that third party web content linked in News results includes hate speech or discriminatory content or content that perpetuates hateful attitudes or harmful stereotypes or otherwise facilitates discrimination towards individuals or groups. • Risk that Copilot Search results or suggested prompts reflect, reinforce, or perpetuate stereotypes, biases, or inequalities in the content generated • Risk that users leverage Bing Image or Video Creator or Copilot Search create discriminatory or hateful content • Risk that Bing Image or Video Creator erases cultural identities through homogenized content creation • Risk that Bing Image or Video Creator might not include or accurately represent diverse abilities of people, accelerating the creation of biased and discriminatory image-based content. • Risk that Bing Image or Video Creator perpetuates stereotypes or produces demeaning representations of individuals with autism or disabilities

	• Risk that Bing Image or video Creator might inadvertently portray genders (including trans, non-binary, etc.) in traditional or offensive roles based on biased training data. This risk can be increased if prompts not accurately or respectfully represent genders in non-English contexts. • Risk that Bing Image or Video Creator is used to produce images that portray certain races, genders, or religions in an unfair light if it's trained on biased data. • Risk that users leverage Bing Image or Video Creator to create discriminatory or hateful content • Risk that Bing Image or Video Creator is used to produce images that degrade or discriminate against specific groups. • Risk that GenAI summaries are based on low authoritative sources and produce content that may be harmful or misleading. • Risk that adversarial user queries or grounding data may bypass the safety features for Copilot Search and produce harmful or misleading content

Inherent Risk

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 "Not Likely."

Bing considered internal metrics and social listening data as key factors in assessing a "Not Likely" level of probability relative to other potential systemic risks on Bing.

The **Inherent Severity** of content and activities that negatively impact Discrimination and Hate on Bing is rated as "High" primarily due to the gravity of risks within this risk area, considering the impact to wellbeing, societal, and economic impacts at the individual-level and societal impacts up to a regional scale. Risks within this risk area, 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 "Not Likely" and Inherent Severity as "High," the **Inherent Risk** associated with Discrimination and Hate on the Bing service is "Moderate."

Maturity of Mitigations

Bing has implemented a robust set of mitigations to address risks related to Discrimination and Hate, including those described in Bing's Approach to User Safety, in Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations 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 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 .14% from September 2024 to April 2025. This means that Bing estimates .14% of search traffic may include 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 highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following risk-specific mitigations focus on measures uniquely tailored to address the identified risks in this area. These are in addition to the broader service-wide mitigations outlined in Bing's Approach to User Safety and the full inventory of Bing-wide protections listed in Appendix II. The mitigations described here include both unique interventions specific to Discrimination and Hate and expansions on previously described cross-platform mitigations with specific applications to Discrimination and Hate Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with bias, hate speech, and discriminatory content. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

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.

Bing has a ***Threat Intelligence*** 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 hateful speech, harassment against protected groups, and other problematic content that could have negative impacts. Bing employs a dedicated expert team that is dedicated to deploying targeted algorithmic interventions to reduce the prevalence of hate speech, and other problematic content in search.

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, RAI and AI-specific risks, hate speech, children and technology, and copyright and trademark infringement, to ensure that its internal policies, practices, and standards are addressing key concerns third-party stakeholders.

External engagements specific to discrimination and hate, Microsoft has engaged groups, such as the Global Project Against Hate and Extremism Project, the Anti-Defamation League (ADL), on issues related to hate speech, with a focus on how content policy and enforcement mitigate the risk of harm in content and GenAI.

Bing provides user transparency and more detailed information on the main parameters it uses for ranking in its ***transparency reports***, specifically 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.

The ***Terms and Conditions for Users,*** found within the Microsoft Service Agreement (MSA), as applied to Discrimination and Hate prohibit using Microsoft services to engage in activity that is harmful to others (e.g., communicating hate speech). The Bing Image Creator Terms of Use further prohibit use of Bing's GenAI features to create or share hateful content; or to create or share content intended to exploit the vulnerabilities of individuals or groups. For example, without limitation, age, disability, or social or economic situations.

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. GenAI 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 "Low."

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Mental and Physical Wellbeing	Not Likely	Critical	Moderate	Defined	Low

Risk Definition

Risk Area	Mental and Physical Wellbeing
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.
Risk Scenario	• Gender-based violence • Individual threats of violence, cyberbullying, harassment, stalking, public order offences, fear, or provocation of violence, and controlling or coercive behavior • Encouraging or assisting suicide or attempted suicide or serious self-harm or self-injury, including eating disorders • Distressing content or activities negatively impacting mental health • Design features encouraging excessive use or creating deeper browsing, potentially leading to negative mental or physical impacts • Content or activities that otherwise negatively impact Mental and Physical Wellbeing
Theoretical Risk Manifestations absent Mitigations	• Risk that search results or auto-suggest features lead users to web results that include gender-based violence. • Risk that search engines return content that could be used to locate content or communities focused on bullying or harassing people in the top search results. • Risk that search results or autosuggest features might lead users to web results that include content that instructs on, promotes, normalizes, praises, or facilitates suicide or self-harm. • Risk that autosuggest features could prompt users to engage more deeply with content promoting or glorifying self-injury. • Risk that search engines return content that promotes unrealistic or unhealthy body standards that could negatively impact mental health, in the top search results. • Risk that search engines return potentially distressing health or other information that can have a negative impact on users' mental health. • Risk that search engines return content that includes graphically violent images or content related to domestic abuse in the top search results. • Risk that Shopping listings show products with logos or graphics that promote include gender based violence. • Risk that News results show content that content that instructs on, promotes, normalizes, praises, or facilitates promotes suicide, self-harm, or causing harm to others.

	• Risk that News results show distressing content or activities that can negatively impact users' mental health. • Risk that Maps locations and listings could lead to offline physical harms. • 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 Advertisements appear on Bing that promote show addictive or harmful products or illegal substances that can negatively impact mental or physical wellbeing. • Risk that Bing Image or Video Creator is used to create content promoting gender-based violence or threatening violence • Risk that Bing Image or Video Creator is used to create material encouraging or depicting self-harm or suicide • Risk that Search or Bing Image or Video Creator results promote unrealistic or unhealthy body standards that could negatively impact mental health

Inherent Risk

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 "Not Likely."

Bing considered internal metrics and social listening data as key factors in assessing a "Not Likely" level of probability relative to other potential systemic risks on Bing.

The **Inherent Severity** of content and activities that negatively impact Mental and Physical Wellbeing is rated as "Critical" primarily due to the gravity of risks within this risk area, 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 risk area, 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 "Not Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with negative impacts to Mental and Physical Wellbeing on the Bing service is "Moderate."

Maturity of Mitigations

Bing has implemented a set of mitigations to address risks related to Mental and Physical Wellbeing, including those described in Bing's Approach to User Safety, in Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations 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 down to "Low."

The Mitigation Maturity for Mental and Physical Wellbeing is rated as "Defined" due to the risks related to a new GenAI product, Bing Video Creator, which introduces potential risk due to its inherent newness and the ability of users to use GenAI products in potentially new and unforeseen ways despite the mitigations currently in place.

While not a proxy for Residual Risk, Bing has measured the DDR for Mental and Physical Wellbeing on Bing Search as an average of .07% from September 2024 to April 2025. This means that Bing estimates .07% of search traffic may include content that may have negative effects related to Mental and Physical Wellbeing.

Nevertheless, Bing continues to invest in developing and enhancing strategies to further mitigate and manage risks related to Mental and Physical Wellbeing. The highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to Mental and Physical Wellbeing and expansions on previously described cross-platform mitigations with specific applications to Mental and Physical Wellbeing. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with gender-based violence, cyber-bullying, and self-harm. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

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 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. 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 potentially related to searching for Child Sexual Abuse Material, directing users to hotlines and report. For Copilot Search, queries related to self-harm or suicide are considered high-risk, and Copilot Search responds to these queries through "thoughtful disengagement" and points users to support options. 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 search interventions are employed to direct users towards high authority content. For responses generated by Copilot Search, there is a high authority approach taken regarding queries relating to eating disorders, to explain the negative effects of certain associated behaviors.

Bing's established "***Right to Be Forgotten***" policies and procedures help reduce risks of negative effects to Mental and Physical Wellbeing by providing reporting tools for users to submit requests that certain online content associated with its name be removed from Bing.

For ***industry engagement***, Microsoft has developed tools and multistakeholder partnerships to combat the rise of misinformation that can pose risks to Public Health or negatively impact Mental or Physical Wellbeing. Partnering with independent third-party organizations empowers Bing product teams to take additional actions to promote more authoritative information.

The ***Terms and Conditions for Users,*** found within the Microsoft Service Agreement (MSA), as applied to Mental and Physical Wellbeing prohibit using Microsoft service to publicly display, generate, or share inappropriate content or material (involving, for example, self-harm) and to engage in activity that is harmful to you or others (e.g., stalking, trying to generate or sharing content that harasses, bullies or threatens others, or advocating violence against others). The Bing Image Creator Terms of Use further prohibit use of Bing's

GenAI features to engage in activity that is harmful to the user, or others; to create or share content that could be used to harass, bully, abuse, threaten, or intimidate other individuals, or otherwise cause harm to individuals, organizations, or society; or to create or share glorification of violence.

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 (NCII), and 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 (NCII). In the EU, Bing respects the Right to be Forgotten.

Risks associated with NCII have shifted in the last year, with increased concerns from the public and regulators about how GenAI exacerbates these risks. In addition to the NCII content, stakeholders are also concerned about access to sites and applications that enable the creation of NCII, so called nudify apps.

Bing made investments in the assessment period to better identify and manage NCII-related content. Bing has a formalized NCII policy that addresses both real and "deepfake" images and is applicable across its ecosystem. Bing also introduced interventions to improve detection and management of NCII, including partnering with StopNCII to detect and remove known NCII based on hashes. Bing also improved the defensive search interventions to better identify prompts and as appropriate, remove potentially harmful content related to NCII and reduce the visibility of nudify apps as low-quality content.

Bing's report a concern form was also updated to capture NCII separately from reporting other privacy issues. These efforts form part of an evolving framework incorporating human review, clear policy enforcement, and low-latency user appeals. Despite these advances, Bing continues to explore ways to enhance its ability to address the complex risks posed by NCII.

As mentioned, Bing has invested in layered defenses for its GenAI capabilities, including upstream filters to detect jailbreak attempts, midstream blocking of high-risk prompts, and downstream controls limiting generation of content related to public figures or harmful subjects. These targeted mitigations, combined with ongoing measurement pipelines and monitoring systems, are regularly refined to ensure effectiveness without compromising user access to information, as referenced above.

Private and Family Life's Residual Risk rating increased from the prior year assessment due to overall increase in the prevalence of NCII internally, as well as in discussions with regulators and civil society, and in online discourse.

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 "Moderate" (See *Changes in Risk Rating* section for further details).

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Private and Family Life	Likely	High	High	Defined	Moderate

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.
Risk Scenarios	• Malicious sharing or exploitation of sensitive or private data and doxing • Disclosure or threats to disclose nonconsensual, private, or intimate imagery or information (NCII) • Privacy intrusions including intrusive marketing practices that exploit personal data • Content or activities that otherwise negatively impact Private and Family Life
Theoretical Risk Manifestations absent Mitigations	• Risk that search engines return content that is doxing individuals or could be used for malicious exploitation. • Risk that search results or autosuggest features lead users to content that includes NCII or leads users, including minors, to synthetic NCII content. • Risk that search engines return content that could be used to find applications, tools, or services for creating synthetic NCII in the top search results. • 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 using or collecting data without explicit consent, potentially leading to unwelcomed privacy intrusions and undermining individuals' autonomy over their personal information. • Risk that private property is featured as a destination or things to do in Travel results. • Risk that bad actors misuse Copilot Search to accelerate the creation of content that could facilitate doxing, exposure of private or intimate information, or identity theft • Risk that Bing Image and Video Creator could be misused to generate synthetic, yet realistic private or intimate images of individuals • Risk that Bing Image and Video Creator could be misused to generate synthetic, yet realistic private or intimate images of individuals

Inherent Risk

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."

Bing considered internal metrics and social listening data as key factors in assessing a “Likely” level of probability relative to other potential systemic risks on Bing.

The **Inherent Severity** of 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 risk area, considering the potential for harm to economic, security, societal, and wellbeing systems at the individual-level with some broader reverberations. Some risks in this risk area, such as posts that contain sensitive data or disclosure of nonconsensual private or intimate images, have potentially irremediable impacts. Some risks in this risk area, such as users’ inadvertent sharing of sensitive personal information (e.g., General Data Protection Regulation (“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.”

Maturity of Mitigations

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 User Safety, in Appendix II: Bing’s Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations 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 "Defined," which brings the **Residual Risk** rating down to "Moderate."

While not a proxy for Residual Risk, Bing has measured the DDR for Private and Family Life on Bing Search as an average of .02% from September 2024 to April 2025. This means that Bing estimates .02% of search traffic may include 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 highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to Private and Family Life and expansions on previously described cross-platform mitigations with specific applications to Private and Family Life. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with NCII, doxing, and privacy intrusions. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

The ***Terms and Conditions for Users*****,** found within the Microsoft Service Agreement (MSA), as applied to Private and Family Life prohibit using Microsoft service to violate or infringe upon the rights of others (e.g., taking photographs or video/audio recordings of others without their consent for any other purpose using any of the Services) and engage in activity that violates the privacy of others. The Bing Image Creator Terms of Use further prohibit use of Bing's GenAI features to engage in activity that violates the privacy of others; to attempt to create or share content that could violate the privacy of others, including disclosure of private information (sometimes known as "doxing"); to conduct facial identification, or identification verification purposes; to input photographs or video/audio recordings of others taken without their consent for the processing of an individual's biometric identifiers or biometric information; or to engage in activity to infer a person's emotional state, age, race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation.

Safety systems in Bing's GenAI features are designed to prevent abuse, by include measures 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 also labels images as generated by Bing Image and Video Creator to limit the possibility of misuse.

Specifically, before any new feature or product is introduced to users, Bing requires it to pass through a robust ***pre-launch risk assessment***. This assessment process is part of Microsoft's broader compliance framework and includes reviews of privacy, security, accessibility, safety, and responsible AI concerns.

These reviews are conducted by professional compliance managers who ensure that products meet Microsoft's standards and legal obligations. Where risks are identified, teams must implement appropriate mitigations before the product or feature can go live. This process ensures that safety, privacy, and integrity are designed into Bing experiences from the start.

For safety, ***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, Bing blurs faces before processing data contained in files provided by the user.

Bing maintains rigorous and principled processes, policies, and procedures for ***content removal and other mechanisms to counter harms***. Central to these processes is a firm commitment to the prevention of CSEAI. Bing proactively scans content to detect known CSEAI using hash-matching technologies such as PhotoDNA and MD5. This scanning occurs before content enters the search index and is used to ensure such material does not appear in results at all, across Core Search and Copilot Search. Bing also proactively scans upload image uploads in Bing Image Creator and Bing Video Creator to prevent known CSEAI from being uploaded. This proactive detection is supplemented by intelligence from trusted partners, including NGOs and law enforcement, and through user reporting channels. When CSEAI is identified, Bing removes it globally and without hesitation. This is a zero-tolerance area where Microsoft's commitment is absolute.

Bing also works to remove adult or sexually explicit images of another person online when shared without that person's consent. This is commonly referred to as nonconsensual intimate imagery (NCII), or "revenge porn". Microsoft considers this to be a gross invasion of personal privacy – whether the content is real or a "deepfake" image that is created using AI or other photo editing tools

Finally, Bing also removes certain other content under global legal standards or Microsoft policies, such as materials containing extremely sensitive personal information that could be exploited for fraud or identity theft. Removals may be prompted by legal orders from governments, user-generated reports, or policy-based decisions following internal investigations. Microsoft applies rigorous standards to each request, including the potential impact on user rights and public interest.

Bing Image Creator uses the ***StopNCII initiative's*** database of hashes for reported and verified NCII content and continuously scans and removes NCII content from Bing's image index.

Bing's ***Report a Concern*** tool was updated during this period to include options for users to report non-consensual intimate imagery, in addition to the exposed personal or private information options previously included.

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.

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 and Activities section above. Bing applies its layered safety mitigations to content that poses risks to Public Health. Bing's ranking algorithms and GenAI safety mitigation systems are designed to protect users from harmful web content and prevent harmful or misleading GenAI 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," and furthermore did not see a change in the scoring for Public Health for this year.

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.
Risk Scenarios	• Promotion of legal but harmful substances, products, or practices • False or misleading health information or misinformation • Health-related disinformation or content that maliciously undermines Public Health institutions or initiatives • Content or activities otherwise negatively impacting Public Health
Theoretical Risk Manifestations absent Mitigations	• Risk that search engines return content that can lead to physical harm, such as dangerous challenges in the top search results. • Risk that search engines return content that includes 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 in the top search results. • Risk that search engines return content that includes low quality information related to health conditions or medical advice in the top search results. • Risk that bad actors intentionally manipulate search results to increase the visibility of health misinformation, including outdated guidance on regulated goods. • Risk that Advertisements on Bing are used to promote legal but harmful substances. • Risk that Bing Shopping shows websites that 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 News shows content containing inaccurate or misleading information about public health issues, medical conditions, treatments, side effects, public health orders or decrees, or other information that could potentially undermine public health. • Risk that Bing Image or Video Creator is misused to generate misleading imagery related to public health issues • Risk that Bing Image or Video Creator results include information or images on legal but harmful substances or practices, or inaccurately represent the effects or consequences of harmful practices or substances • Risk that Bing Image or Video Creator is used to create images that glamorize the use of harmful but legal substances • Risk that GenAI features generate ungrounded results or mischaracterize cited web links that leads to misinformation about health. • Risk that GenAI features are based on low authoritative sources and produce content that may be harmful or misleading.

Inherent Risk

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."

Bing considered internal metrics and social listening data as key factors in assessing a "Not Likely" level of probability relative to other potential systemic risks on Bing.

The assessment team concludes as time passes since the COVID-19 pandemic, public discourse related to Public Health risks has decreased on Bing's services, as indicated by social listening data, survey data, and internal metrics.

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 risk area, 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 risk area, 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."

Maturity of Mitigations

Bing has implemented a robust set of mitigations to address risks related to Public Health, including those described in Bing's Approach to User Safety, in Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations 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 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 .04% from September 2024 to April 2025. This means that Bing estimates .04% of search traffic may include 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 highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to Public Health and expansions on previously described cross-platform mitigations with specific applications to Public Health. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with health misinformation. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

Health responses are only generated using ***grounding data from trusted health domains*** which have been vetted by the Bing Health team.

Bing has a ***Threat Intelligence*** 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 Public Health, and other problematic content that could have negative impacts.

For ***authoritative content and crisis hubs***, Bing actively directs users to trusted sources through several tools. For high-impact events such as various EU elections, Bing highlights answers and PSAs that point to expert-vetted information. During events like Public Health crises, Bing would apply defensive search

interventions such as demoting low authority results in ranking on queries that may lead users to data voids and misleading content informed by threat intelligence.

For ***industry engagement***, Microsoft has developed tools and multistakeholder partnerships to combat the rise of misinformation that can pose risks to Public Health or negatively impact physical or mental wellbeing. Partnering with independent third-party organizations empowers Bing product teams to take additional actions to promote more authoritative information.

The ***Terms and Conditions for Users***, found within the Bing Image Creator Terms of Use, as applied to Public Health, prohibit use of the service to engage in activity that is false, or misleading; to create or share content that could mislead or deceive others, including for example creation of disinformation; or to attempt to mislead others about whether the Creations were generated by the service.

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 GenAI safety mitigation systems are designed to protect users from harmful web content and prevent harmful or misleading GenAI 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 Microsoft's 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 GenAI.

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 "Low;" which, as mentioned above in the *Changes in Risk Rating* section, experienced a Residual Risk Rating decrease from a "Moderate" level.

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

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.
Risk Scenario	• Content or conduct that promotes, encourages, or facilitates violent extremist or terrorist activities; violent extremist or terrorist recruitment, funding, training, or incitement; violent extremist or terrorist imagery or content; or cyberterrorism. • Organized crime, cyberattacks, coordination or cooperation to inflict physical harm or death, or compromise public infrastructure and security • Misinformation related to crisis events • Disinformation undermining public security institutions or initiatives or related to crisis events • Content or activities that otherwise negatively impact public security

Theoretical Risk Manifestations absent Mitigations	• Risk that search engines return terrorist recruitment content. • Risk that search engines return terrorist- and violent extremist-operated websites that disseminate propaganda or recruit members. • Risk that search engines return content that promotes violent acts. • Risk that search engines return content that includes information on enabling users to build harmful weapons or devices for use in a terrorist or violent extremist attack. • Risk that search engines return content where online groups or communities engage in ideation and planning for a mass-violence attack. • Risk that Autosuggest features lead users to web results that instruct on how to work with others to avoid detection while planning violent attacks. • Risk that search engines return content that is inaccurate, misleading information, harmful misinformation, related to crisis events or issues of public security in the top search results. • Risk that search engines return content that includes disinformation related to crisis events or issues of public security in the top search results. • Risk that activities on Shopping are used to collect funds used to finance terrorist organizations. • Risk that Shopping is used to assemble harmful weapons or devices for use in a terrorist or violent extremist attack • Risk that Maps and Travel show locations or rental listings that are targeted for terrorist training, funding, or support for terrorist acts. • Risk that News results include inaccurate, misleading information or issues of general public security in EU member states and around the world. • Risk that News results include disinformation related to crisis events or issues of general public security • Risk that Maps' aerial and street level imaging of sensitive locations, including military bases, may expose national security assets. • Risk that Bing Image or Video Creator is used to generate images that contribute to public security risks • Risk that Bing Image or Video Creator is used to create Terrorist propaganda or imagery • Risk of Bing Image or Video Creator being used to create realistic threat scenarios or emergency event simulations that could cause public panic or hinder emergency response efforts • Risk that Bing Image or Video Creator is used to generate infographics or images that perpetuate a misrepresentation of public security events • Risk that Bing Image or Video Creator is used to generate visual propaganda that is used to disseminate mis- or disinformation related to crisis events • Risk that GenAI summaries generate ungrounded results or mischaracterize cited web links that leads to misinformation about public safety incidents. • Risk that GenAI summaries are based on low authoritative sources and produce content that may be harmful or misleading. • Risk that adversarial user queries or grounding data may bypass the safety features for Copilot Search and produce harmful or misleading content

Inherent Risk

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 "Not Likely."

Bing considered internal metrics and social listening data as key factors in assessing a “Not Likely” level of probability relative to other potential systemic risks on Bing.

The **Inherent Severity** of content and activities that negatively impact Public Security on Bing is rated as "Critical," primarily due to the gravity of risks within this risk area, 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 risk area is irremediable.

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

Maturity of Mitigations

Bing has implemented a robust set of mitigations to address risks related to Public Security, including those described in Bing's Approach to User Safety, in Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations 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 down to "Low."

While not a proxy for Residual Risk, Bing has measured the DDR for Public Security on Bing Search as an average of .07% from September 2024 to April 2025. This means that Bing estimates .07% of search traffic may include content that may have negative effects related to Public Security.

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

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to Public Security and expansions on previously described cross-platform mitigations with specific applications to Public Security. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with terrorism and crisis events. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

Bing takes a multi-stakeholder approach by having ***External Partnerships related to Public Security***. For example, 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 services, Bing has access to the GIFCT's Incident Response processes. This allows Bing to quickly become aware of, assess, and address potential content circulating online resulting from a terrorist or violent extremist event.

Bing participates in the Advisory Network (a multistakeholder advisory group) of the Freedom Online Coalition, a coalition of 37 governments working to advance Internet freedom.

Bing also 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.

The ***Terms and Conditions for Users*****,** found within the Microsoft Service Agreement (MSA), as applied to Public Security, prohibit using Microsoft service to engage in activity that is harmful to others (e.g., posting terrorist or violent extremist content). The Bing Image Creator Terms of Use further prohibit use of Bing's

GenAI features to create or share terrorism and violent extremist content; to engage in activity that is false or misleading; to attempt to create or share content that could mislead or deceive others, including for example creation of disinformation; or to attempt to mislead others about whether the Creations were generated by the service.

Rights and Protection of Minors

Risks related to Rights and Protection of Minors include exposure of minors to harmful behavioral activities and content and collection of minors' data without parental consent. Bing is committed to prioritizing the rights and safety of minors, aligning its product designs with the principles of the United Nations Convention on the Rights of the Child ("UNCRC"). 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 enables parents to manage and track use of applications, including Bing.

Bing does not require users to authenticate to access search. This is standard practice across search engines and is of the utmost importance given the role that search engines play in enabling users to freely access 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 GenAI 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 the Rights and Protection of Minors, creative GenAI services for Bing Image and Video Creator features are available only to authenticated users over the age of 13 (or over the age of parental consent if higher for their jurisdiction). Full functionality for Bing Image and Video Creator is disabled for all unauthenticated users.

Copilot Search defaults to Core Search results for any complex or sensitive queries, which already has its own layered restrictions, and therefore does not require the same age-based limitations as Bing Image and Video Creator. As a result, Copilot Search is available to all signed-in users without explicit age restriction.

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 "Low."

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Rights and Protections of Minors	Not Likely	Critical	Moderate	Defined	Low

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 of minors or respect for the rights of the child occur on the service.
Risk Scenario	• Child sexual exploitation and abuse, grooming, and trafficking of minors • Collection of children's data without parental consent, or targeting children with personalized advertisements

	• Exposure of children to harmful content or activities, including gambling, exploitation, crime, graphic violence, human gore, terrorism, or violent extremism • Discrimination, cyberbullying, and harassment of minors • Promotion of self-harm or self-injury for minors • Unnecessary restriction from access to information • Content or activities that otherwise negatively impact the rights or protection of minors
Theoretical Risk Manifestations absent Mitigations	• Risk that search engines return content that instructs on, normalizes, promotes, or facilitates sexual exploitation, grooming, and abuse of a child or pedophilia. • Risk that algorithmic or ranking systems inadvertently elevate results that aid human traffickers in recruiting by promoting false job postings that appeal to minors such as modeling or acting. • Risk that search engines return content that contains minors' personal data which can be used by adult users to locate, bond, and meet with children or used by abusers to target and groom minors. • Risk that Bing's data processes harm minors or collect its data without parental consent. • Risk that search results or autosuggest features lead minor users to websites that normalize harmful third-party content, such as hate speech, extremist content, violent or graphic material, adult content, information promoting or soliciting illegal and regulated goods, self-harm or content inciting violence against others. • Risk that search engines return content that shows risky behaviors to minors, such as dangerous challenges in the top search results. • Risk that auto-suggested searches expose minors to content that perpetuates discrimination and cyberbullying. • Risk that search engines return content to minors that normalizes eating disorders in the top search results. • Risk that Bing's safety systems for search unduly limit minors' access to information. • Risk that Advertisements on Bing expose minors to harmful activities targeted for child sexual exploitation and abuse • Risk that Shopping, Maps, News, and Ads may harm minor users by collecting minor's data without parental consent and target children minors with personalized content. • Risk that Advertisements on Bing target minors. • Risk that Shopping, News, and Ads results 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, including: Weapons and firearms, recreational drugs and paraphernalia, prescription drugs and pharmaceuticals, gambling, alcohol and tobacco products. • Risk that those age-inappropriate activities are recommended for children and/or families in Travel guides, or on the Maps. • Risk that Shopping or News could expose minors to products or content that perpetuate discrimination. • Risk that Bing's safety systems for Ancillary Features unduly limit minors' access to information. • Risk that images of children which do not conceal their privacy are shown on the "Travel the World Virtually" or in the Maps features. • Risk that Bing Image or Video Creator generates harmful content inappropriate of/for minors • Risk that Bing Image or Video Creator minor user prompts could be stored or collected without parental consent. • Risk that Bing Image or Video Creator is used to create imagery that glorifies and promotes self-harm for minors. • Risk that Bing Image or Video Creator promotes risky behaviors among minors, such as dangerous challenges or trends, through compelling images

Inherent Risk

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 "Not Likely."

Bing considered internal metrics and social listening data as key factors in assessing a “Not Likely” level of probability relative to other potential systemic risks on Bing.

It is important to note that the Bing service is not known to be used by a large number of minors (roughly 3% 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 GenAI 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 risk area, the vulnerable nature of the population impacted, the irremediability of impact of certain risks, and the potential for societal impact broader than the individual scale for risks impacting youth. Not all risks within this risk area are assessed at the highest level of severity, namely collection, processing, and use of minors’ data without parental consent.

With the rating for Inherent Probability as “Not 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 “Moderate.”

Maturity of Mitigations

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 User Safety, in Appendix II: Bing’s Digital Safety Risk and Compliance Framework and Mitigations, and those risk-specific mitigations 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 down to “Low.”

The Mitigation Maturity rating of “Defined” reflects the presence of numerous sophisticated, risk-specific mitigations. However, this rating also acknowledges certain limitations. One factor influencing the rating is that Copilot Search does not restrict minors from accessing the service, meaning Copilot Search services should be considered as part of the Rights and Protection of Minors Risk Area. Another factor is the recent removal of the risk of CSAM from the Rights and Protection of Minors risk area, with associated risks and mitigations now realigned under the Illegal Content risk area. These aspects, while not increasing residual risk materially, are taken into account in assessing the in scope mitigations.

While not a proxy for Residual Risk, Bing has measured the DDR for Rights and Protection of Minors on Bing Search as an average of .54% from September 2024 to April 2025. This means that Bing estimates .54% of search traffic may include content that may have negative effects related to Rights and Protection of Minors.

Nevertheless, Bing continues to invest in developing and enhancing strategies to further mitigate and manage risks related to Rights and Protection of Minors. The highlights of the key mitigations currently implemented are described below.

Risk Specific Mitigations

The following section highlights both how broader service-wide mitigations described in Bing's Approach to User Safety and inventoried in Appendix II, apply for this risk area, and include mitigations that are uniquely tailored to address the identified risks in this area.

The mitigations described here include both unique interventions specific to the Rights and Protection of Minors and expansions on previously described cross-platform mitigations with specific applications to the Rights and Protection of Minors. Where applicable, Bing's broader infrastructure has been adapted to address the particular risks associated with gender-based violence, cyber-bullying, and self-harm. This layered approach ensures both platform-wide consistency and harm-specific responsiveness.

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.

In response to user search queries for CSAM, Bing displays ***in-product indicators*** and PSAs to help protect minors, similar to those regarding negative searches regarding suicide, stating that CSAM is illegal and depending on the query may not return search results; Bing also includes links to report these illegal materials or receive anonymous support, aimed at those who are victims of CSAM.

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 setting age restrictions under the Microsoft Services Agreement ("MSA"), developing age-appropriate privacy communications, ensuring Bing features are accessible and understandable to teen users, as well as enforcing strict advertising policies through Microsoft Advertising. Additionally, Bing's commitment to privacy protection for minors is evident through practices such as 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.

Bing Image and Video Creator is 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.

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. 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 its policies and reserves the right to remove those that are non-compliant.

For ***pre-launch risk assessments***, 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.

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. 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.

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.

Bing has ***externally 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.

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, 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 facilitated an internal meeting with civil society partners, to explore two key topics,1) state privacy legislation, and 2) minor data and AI. Microsoft presented some draft guiding principles and foundational capabilities for developing AI products and services for young people. Feedback from external engagements 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.

Microsoft also offers a child-friendly version of its ***Microsoft 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.

Bing's ***Content Moderation Ops team carries out training*** for content reviewers 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.

The ***Terms and Conditions for Users***, found within the Microsoft Service Agreement (MSA), as applied to the Right and Protection of Minors prohibit using Microsoft service to engage in any activity that exploits, harms, or threatens to harm children.

Conclusion

Conclusion of Assessment

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 (as outlined in the *Introduction, Bing Overview and Risk Profile* and *Bing Risk Results* sections), 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 (See the *Risk Specific Mitigation Section* for *Freedom of Expression and Information above)*. 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, Bing works to include additional contextual information to search results. These interventions may include for example, answers, PSAs, site-level warnings, indicators of content provenance, 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 0.22% 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 (>95%) 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 responds to search queries with the highest authority, relevant content, and grounds GenAI responses in the same principles.

The lack of user-generated content and user-to-user interaction on the service

As previously noted within the *Core Search Features* section, since Bing generally does not allow users the ability to post content or communicate directly with other users, this further limits Bing's risk exposure as to many other systemic risk areas common to online services, 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 (See *User Base and Market* section above). 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.

Collaboration with key stakeholders to inform risk measurement, mitigation, and broader safety considerations

As previously mentioned in the *Research and Partnership* and *Risk Specific Mitigation* sections, 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, RAI, 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 importance of continuing to assess and manage 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.

As beforementioned, after evaluating the purpose and design of Bing's core search, GenAI, 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 service.

Progress on Enhanced Mitigations from Bing's Last Report

In the 2024 Systemic Risk Assessment report, Bing outlined six primary areas for enhancement of mitigations during the initial assessment period. Bing has made enhancements in each identified area as described below.

Governance and Organizational Alignment

To support durable systemic risk mitigation, Bing has reinforced its internal governance structures. This included formalizing safety governance across the Microsoft AI organization. Roles and responsibilities across Bing and associated product teams have been clarified and documented within the SRA Playbook to ensure accountability where systemic risk intersects multiple service areas. This structural clarity supports proactive ownership and streamlines escalation paths in the context of compliance obligations under Article 35(f).

Risk Detection and Monitoring Infrastructure

Bing continued to scale its investment in systemic risk detection through enhanced classifier development and risk mapping. Core efforts included enhanced alignment of internal classifiers to systemic risk areas, and classifier development to better track evolving threats. These classifiers have been integrated into the DDR pipeline and support ongoing monitoring. Complementing these efforts was the development of dashboards for risk monitoring, operational feedback loops, and compliance traceability.

Documentation and Incident Response

A critical focus has been placed on improving documentation standards across the systemic risk lifecycle. This included mapping how risks manifest in features, detailing mitigations in place, and measuring their effectiveness. A new team member was onboarded to centralize internal policy documentation. Incident response protocols were standardized. This work supports Articles 34 and 35 by ensuring procedural transparency and organizational learning and addresses current inconsistencies in documenting responses to content-related or operational incidents.

User Reporting and Notice-and-Action Mechanisms

Bing is refining and expanding its existing user reporting mechanisms to better capture nuanced harms and improve mapping between user-submitted concerns and specific product features or systemic risk areas. These updates build on existing notice systems and aim to enhance usability, support more detailed internal risk analysis, and provide richer insight into the user experience of risk. While implementation depends on migration to Microsoft's DigiTS RaC system, the intent is to support richer insights into user experience of risk. These changes do not replace or imply the inadequacy of prior systems but serve to evolve and strengthen reporting capabilities over time.

External Engagement and Consultation

Consistent with Recital 90 Bing is continuing to engage with civil society groups and issue-area specialists with the goal to integrate diverse perspectives into risk assessment and mitigation design, especially in high-impact areas such as youth safety, misinformation, or hate speech. A tracking process is in place to capture expert input and ensure it informs the refinement of risk mitigation measures. This supports more evidence-based and inclusive risk management practices.

Feature-Level Risk Quantification and Crisis Response

Bing is building infrastructure to calculate MAUs segmented by key features, login state, and age group. This supports a more precise assessment of probability and potential impact for systemic risks as required under

Article 34. Separately, crisis response capabilities are being strengthened, particularly for Maps. For example, during active crises, location-based recommendations will be suppressed in the Local Guide carousel to prevent misinformation or safety risks. These targeted interventions ensure that Bing can respond swiftly and appropriately in emergent situations.

Looking Ahead to the Coming Year

Bing has made investments in digital safety and risk mitigation throughout this reporting period, and it looks forward to continuing to build upon and improve its approach to mitigating systemic risks in the EU and adapting to new challenges ahead.

The following areas have been identified for heightened focus for the next assessment 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 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.

Enhancing governance and pre-launch review.

With the rapid pace of experimentation, learning, and innovation in search, including Bing, the team continues to leverage existing internal governance structures, while also improving those to increase scale, agility, and efficiency, and to more easily demonstrate compliance with regulation and corporate policies.

Developing and refining policies to align user experience with Bing's Trustworthy Search Principles.

To further refine Bing's Trustworthy Search Principles as outlined in *Bing's Values and Commitments* section, Bing has developed a policy framework and process that enables more cohesion across the different experiences that could result from users' queries. Leveraging this framework, Bing is also engaging earlier and more predictably with teams on product changes. In addition, Bing continues to build out policy documentation and resources to more closely align with operations, creating a flywheel of continuous learning and improvement in its management of systemic risks.

Continuing investments in targeted interventions to better manage systemic risks.

Bing continually invests in refining and enhancing its limited, targeted interventions designed to manage systemic risks effectively. For example, Bing is focused on improving threat intelligence capabilities to better detect and anticipate emerging risks. Bing is also consistently refining query and result classifiers to improve accuracy in identifying potentially harmful results and to ensure safer, more reliable information surfaces.

Refining documentation, resources, and metrics to enable greater data-driven improvement.

Bing continues to evaluate and, as appropriate, update user-facing resources, such as reporting mechanisms, as well as internal metrics to enable more nuanced and context-aware data analysis. This increased granularity in data enables Bing to identify opportunities to improve the management of systemic risks.

Improving children's safety.

As a part of Bing's ongoing commitment to user safety, and bolstered by regulatory changes, Bing is updating the protections provided for users that are in Strict or Moderate mode in SafeSearch in Bing. We believe these changes will create a better experience for those users, mitigating the risk that they might unexpectedly

encounter harmful, offensive, or illegal content, while not limiting users' ability to see or access sensitive content with SafeSearch in Off or Moderate.

Risk assessment and mitigation is and should be a continuously evolving process as threats and risks evolve and as Bing continues to engage in innovation and experimentation with respect to Search. Bing is committed to continuously monitoring the effectiveness of the implemented risk mitigation to address issues as they arise and continue to refine mitigations.

Appendix I: Detailed Systemic Risk Assessment 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 Definition
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, impacting individuals and groups based on race, ethnicity, religion, gender, sexual orientation, or other protected traits, or reflecting, reinforcing, or perpetuating harmful stereotypes, biases, or inequalities.
4	**Freedom of Expression and Information**	Risk that content or activities with actual or foreseeable negative effects on the fundamental right 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 Mental and Physical Wellbeing 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 GenAI 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.

- **Stakeholder validation of mitigations:** The Bing Risk Assessment team deployed an information collection sheet with associated questions to each Bing product team to systematically identify changes to Bing products and services since August 2024 to be described in the 2025 Systemic Risk Assessment.
- **Authoritative sources**: The Bing Risk Assessment team reviewed publicly available sources considered reliable due to their expertise and reputation, including the European Commission, CERRE and other sources such as such as civil society and other non-governmental organizations, on the severity of systemic risks.

- **Policies and publications:** The Bing Risk Assessment team reviewed relevant internal policies and publications, 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.
- **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 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.
- **Open-source data on public discourse:** 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.
- **Product and feature material changes:** The Bing Risk Assessment team evaluated product and feature material changes that occurred over the past year.
- **DSA Risk and Control Matrix:** The Bing Risk Assessment team incorporated controls outlined in the DSA Risk and Control Matrix and potential mitigations where relevant to systemic risks.

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 rating.

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:

- Open-source data on public discourse
- Transparency report metrics
- Internal metrics
- Internal consultations
- External consultations
- Stakeholder validation of mitigations

Figure 14: Probability Rating Scale

Description	Weight	Score	Rating
The risk event or circumstance is relatively certain to occur without considering mitigations in place	>=243	5	Expected
The risk event or circumstance is highly likely to occur without considering mitigations in place	193 – 243	4	Highly Likely
The risk event or circumstance is likely to occur without considering mitigations in place	143 – 193	3	Likely
The risk event or circumstance occurring is possible but not likely without considering mitigations in place	93 – 143	2	Not Likely
The risk event or circumstance is only remotely probable without considering mitigations in place	<93	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 15: 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	Moderate

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 16: Severity Rating Scale

Description	Weight		Score	Rating
	Min	Max		
Impact that could cause critical, irremediable harm, damage, or loss	145	270	5	Critical
Impact that could cause significant irremediable harm, damage, or loss	95	145	4	High
Impact that could cause some harm or disruption but is generally manageable or remediable	75	95	3	Moderate
Impact that could cause limited harm or disruption	71	75	2	Low
Impact that has little or no consequence	70	71	1	Minimal

Inherent Risk

Inherent Risk is a multiplication of Inherent Probability by Inherent Severity to identify the level of risk absent mitigations.

Figure 17: Inherent Risk Rating Scale

Description	Score (>=)	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.	0	2	Minimal

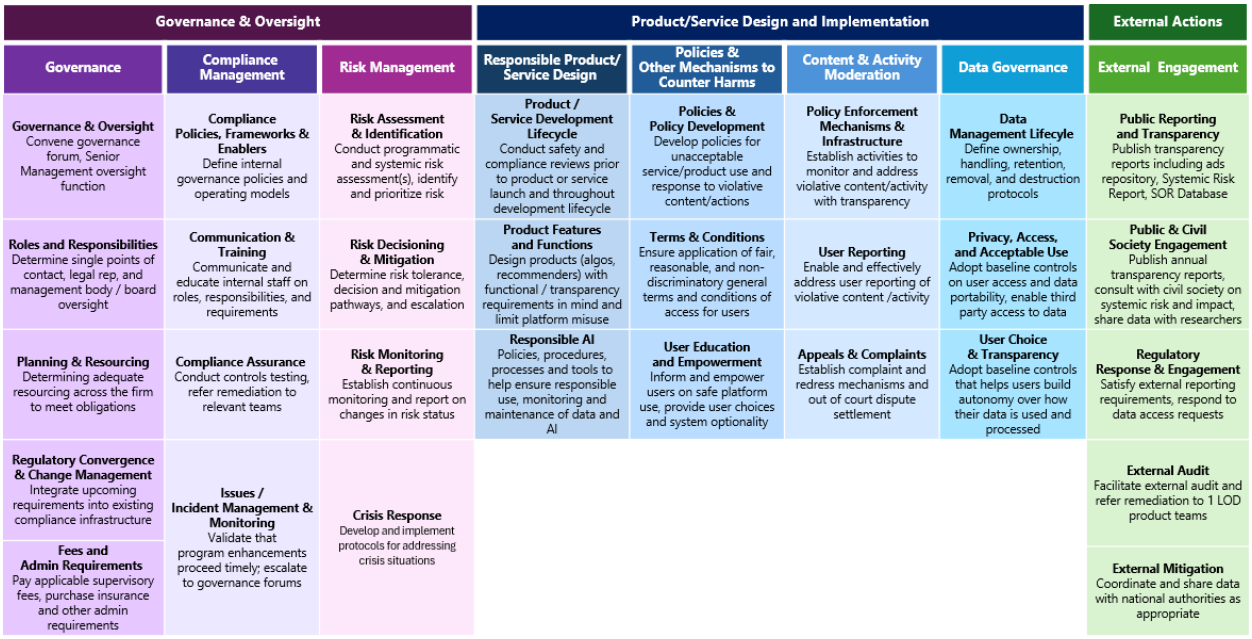
Maturity of Mitigations

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 Bing Risk and Compliance framework to ensure coverage across industry-standard mitigation practices. Utilizing the Bing Risk and Compliance Framework enables Bing to set standards and enact best practices from peers across the technology industry, enhanced 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.

Bing stakeholders complete product specific information collection sheets to identify implemented best practices within the listed Bing Risk and Compliance Framework 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 18: Bing Risk and Compliance Framework Best Practices Screenshot



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 rating higher than Managed maturity.

Figure 19: 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:

- Stakeholder validation of mitigations.
- Guidance from authoritative sources, including the European Commission, on the severity of systemic risks.
- Collection of Bing's published policies and official communications.
- Bing internal stakeholder summarized policies and procedures.
- Notes from internal consultations.
- Notes from external consultations.
- Bing internal metrics.
- Transparency Reporting metrics.
- Open-source data on public discourse related to Bing and systemic risk areas.
- Bing product and feature material changes.
- DSA risk and control matrix.

Figure 20: Mitigation Maturity Scale

Description	Risk Coverage	Framework Coverage	Final Score	Rating
A rating of Ad Hoc is assigned when execution of best practices is incomplete, informal, or inconsistent.	0	N/A	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.	5	N/A	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. A risk can only receive a Defined coverage rating if there are mitigations in place on each product to address all risks relevant to that product.	10	N/A	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. A risk can only receive a Managed rating if, in addition to having full risk coverage, the mitigations in place cover 50% of framework categories.	15	50%	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. A risk can only receive an Optimized rating if, in addition to having full risk coverage, the mitigations in place cover 100% of framework categories.	N/A	100%	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 21: Residual Risk Rating Scale

Description	Score (>)	Score (<=)	Score (<=)
A risk that is relatively certain to occur would have a severe impact if it occurred and requires immediate action to manage or mitigate.	15.1	<=	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.	10.1	15	High
A risk that is likely occur, would have a noticeable impact if it occurred, and requires some action to manage or mitigate.	5.1	10	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	5.1	Low
A risk that is only remotely probable, would have a minimal impact if it occurred, and may not require any action to manage or mitigate.	0	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 22: Systemic Risk Scoring Equation Screenshot



Appendix II: Bing's Digital Safety Risk and Compliance Framework and Mitigations

Below is a summary of Bing's mitigations that span multiple systemic risk areas. The mitigations below have been categorized across the Bing Safety Funnel. Although mitigations may address more than one component of the Bing Safety Funnel best practice, mitigations have been categorized based on the Bing Safety Funnel component it primarily addresses.

Product Quality

Terms and Conditions – Generative Experiences: GenAI Terms and Conditions governing the use of Bing's GenAI features are outlined in the Bing Image Creator Terms of Use. These provisions prohibit activities that could negatively impact users, such as engaging in harmful, fraudulent, or illegal activities, violating privacy, or creating inappropriate content. Users who violate these terms may face limitations on their access to the services.

Terms and Conditions – Microsoft Services Agreement: Bing publishes and enforces its content standards through terms and conditions, specifically through the Microsoft Services Agreement and the Bing Image Creator Terms of Use which, among other conditions, prohibit the generation or sharing of inappropriate or harmful content. These policies apply globally and are available in local languages. Bing reserves the right to suspend or ban users who attempt to misuse the service. For example, entering prompts designed to bypass safety systems is not permitted. These terms help ensure that Bing remains a safe and transparent service for all users.

Content and Activity Moderation:

Safety Measures – Search Suggestions: Bing has implemented measures aimed at curbing the generation of search suggestions that might tarnish users' fundamental rights. 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.

SafeSearch: Bing also offers users direct control over what they see through SafeSearch, a feature that uses machine learning to automatically detect and filter adult and explicit content from search results. This feature is especially important in educational and household settings where additional safeguards may be required.

SafeSearch settings can be locked to prevent tampering and ensure consistent application of filters. Bing also provides a mechanism for users to report inappropriate content or suggestions. These reports feed into the continuous improvement of the underlying safety systems, helping Bing respond quickly to emerging issues.

Content Provenance & Watermarks: Bing invests in helping users identify AI-generated content. 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.

Metaprompting: To guide GenAI models in generating responses that align with Microsoft's AI Principles and user expectations, Bing employs a technique known as "metaprompting." Metaprompting involves providing instructions to the AI model to influence its behavior, ensuring that the system operates responsibly and ethically.

Safety by Design – GenAI Features: The development and deployment of GenAI features at Bing are guided by safety processes that ensure adherence to high standards of responsibility and emphasize safety by design for GenAI features. These processes include red team testing, expanded reactive social listening systems, prominent reporting functionality, and operations and incident response mechanisms. These measures help identify vulnerabilities and inform the development of targeted mitigations.

Content Moderation – 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 sudden images. Bing Real Estate automatically detects fraudulent content and continues to improve Real Estate's fraud detection capabilities as new methods for posting harmful content arise. This scanning is complemented by allowing users to report listings that they find inappropriate or fraudulent.

Data Governance:

GenAI Disclosures: Copilot Search and Bing Image and Video Creator 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 Search 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.

Citation Links: Copilot Search responses include citation links, and the corresponding web results are displayed at the bottom of the results. This practice allows users to verify the information and assess the credibility of the sources.

User Privacy Choices: Respecting user privacy choices is a core Microsoft commitment, and as part of Product Quality, Bing provides users with control over how their data is collected, used, and personalized. Through the Microsoft Privacy Dashboard, users can view, delete, or export their personal data, including search history. They can also adjust advertising and content personalization settings. For example, if a user frequently searches for travel content, Bing might personalize their Microsoft News feed with travel-related stories. These personalization features are optional and can be disabled. In the EU, Bing also requires users to opt in before using cookies or similar technologies that collect data for personalization.

Policies and Other Mechanisms:

Prioritizing user fundamental rights to search: As a search engine, Bing is committed to supporting and prioritizing users' fundamental rights and their rights to access and seek information. Bing does not broadly remove content from its index. Instead, it focuses on minimizing unintended exposure through ranking and targeted filtering. Where content is removed from search, Bing is transparent by notifying users through a notice on the search engine results page and publishing transparency reports.

Classifiers: Bing deploys specialized mitigations in its GenAI features, such as "classifiers," to reduce the risk that harmful content might be generated by Bing. These classifiers analyze text, images, and videos to identify potentially harmful content in search queries and generated responses. By applying AI-based classifiers and content filters, Bing can flag and prevent the production of harmful content, including in

features such as Copilot Search and Bing Image and Video Creator. For instance, classifiers have been implemented to mitigate the risks of generated content related to public figures and to prevent the generation of low-authority content.

Terms & Conditions – Webmasters: Bing is transparent about its policies and actions with respect to webmaster content and makes these documents available in local languages. Bing's Webmaster Guidelines are accessible via the link.

Responsible Product/Service Development:

GenAI – Grounding in High-Ranking Search Results: Copilot Search responses are generally grounded in high-ranking search results, a process known as "grounding." This approach involves providing data to a large language model (LLM) to improve its ability to generate accurate, relevant, and updated outputs. By centering responses on high-authority content from the web, Bing aims to prevent the generation of content containing private information hosted on low-authority sites. Additionally, providing links to source websites allows users to evaluate the credibility of the information presented.

Continuous Improvement of Ranking Algorithms: Bing's algorithms are the first and strongest defense in delivering safe, authoritative content. These systems are designed to match user queries with the most relevant third-party web pages in Bing's index, elevating trusted sources and reducing visibility of unreliable or manipulated content. Rather than removing content outright, Bing relies on continuous improvement of ranking algorithms to ensure that harmful content is not surfaced prominently, particularly when users have not explicitly searched for it. To support this, Bing invests continuously in refining how its crawlers and algorithms detect quality content. These improvements are applied across all markets and languages where Bing operates. Teams regularly assess performance through key metrics, including internal safety indicators, and use these results to inform updates. In addition to technical signals, Bing incorporates feedback from users, Microsoft's policy teams, regulators, and civil society organizations to ensure its ranking decisions reflect evolving standards of trust and accuracy.

Pre-Launch Risk Assessment: Specifically, before any new feature or product is introduced to users, Bing requires it to pass through a robust pre-launch risk assessment. This assessment process is part of Microsoft's broader compliance framework and includes reviews of privacy, security, accessibility, safety, and responsible AI concerns.

These reviews are conducted by professional compliance managers who ensure that products meet Microsoft's standards and legal obligations. Where risks are identified, teams must implement appropriate mitigations before the product or feature can go live. This process ensures that safety, privacy, and integrity are designed into Bing experiences from the start.

Data Protection Impact Assessment ("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 reevaluation 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.

Cultural Sensitivity Safeguards – Maps: Bing 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.

Risk Management:

Safety Measures – Prioritizing High-Authority Content in Search: Bing's commitment to Trustworthy Search Principles means prioritizing high-authority content in search results, especially in areas where inaccurate or manipulated content may pose risks. Rather than removing content outright, Bing relies on ranking algorithms to ensure that harmful content is not surfaced prominently, particularly when users have not explicitly searched for it.

To support this, Bing uses a Quality and Credibility ("QC") score, which includes external signals that help identify authoritative sources. These signals help elevate trusted results while demoting pages that may be misleading or exploitative. Bing also uses ranking systems to defend against manipulation, including spam or search engine optimization abuse that could otherwise distort results.

These ranking systems are continuously measured and refined to ensure they reflect user expectations, policy goals, and emerging risks. The result is a more accurate and safer search experience that remains grounded in user intent.

Product Safety

Compliance Management

Internal Training: 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 receive additional training and regular communications from expert teams in their areas of focus, such as for digital safety, privacy, or RAI. Staff with specific safety responsibilities must complete at least one digital safety specific training per year.

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.

Content and Activity Moderation

Safeguarding the rights and protection of minors: 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 setting age restrictions under the MSA, developing age-appropriate privacy communications, ensuring Bing features are accessible and understandable to teen users, as well as enforcing strict advertising policies through Microsoft Advertising. 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.

Blocklists: Bing Image and Video Creator uses targeted safeguards to limit the generation of inappropriate or potentially harmful content. A key intervention is the use of intelligent blocklists—filters designed to detect

and restrict specific user inputs that contain sensitive or risky names, terms, or phrases. These inputs are sorted into "hard" or "soft" blocklists, depending on the level of sensitivity and risk associated with them.

The development and refinement of these blocklists is a collaborative process involving internal product and safety teams, insights from jailbreak researchers, red teams, and specialized legal counsel. For example, during election periods, Bing consults with elections counsel to ensure the blocklists reflect heightened sensitivities around electoral misinformation or manipulation. This multi-disciplinary approach allows Bing to tailor safeguards to real-world risks in a dynamic and proactive manner.

Age Restrictions for GenAI Features: Bing has certain age restrictions for GenAI features. While some, limited functionality in Bing Image and Video Creator is available to all unauthenticated users, full functionality is restricted to authenticated users who meet the appropriate age thresholds. Users must be signed in with a Microsoft account that reflects an age of 13 or older, or a higher age in jurisdictions where parental consent laws apply, to access full functionality, including multiple turns on the same content. If a user's account indicates they are below the required age, they are prevented from accessing Bing Image and Video Creator, even the limited experiences available to unauthenticated users. This age-based access control helps protect younger users from exposure to content or interactions that may not be appropriate for development.

Prevention of Targeted Advertising to 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 its policies and reserves the right to remove those that are non-compliant.

Privacy Training: Bing has implemented a multifaceted approach to product enforcement to protect user privacy. 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.

Data Governance

Safeguarding User Data: 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. 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. The Visual Search 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. Access to most Bing Search history data is limited to Bing employees. 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.

Safety Measures – Privacy: Bing's terms of use governing activity on its GenAI features prohibit the use of

the service to violate others' privacy. Safety systems in GenAI 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 Bing Image Creator to limit the possibility of misuse.

Data Retention and Deletion Policies: Bing 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.

Unauthenticated users can control their search history through Bing's in-product controls using Bing's settings page.

External Engagement

Verified Project: Microsoft has extended its support to the Verified Project in response to disinformation campaigns targeting Slovakia. Launched in the fall of 2023, Verified 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, Verified is able to recognize manipulative elements in the text.

Governance

Microsoft Threat Assessment Center: Bing is actively engaged in enhancing public security through the collaboration of its internal teams, such as the 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.

Policies & other Mechanisms to Counter Harm

Special Authoritative Elections Answer in Search: Bing launches special informational and Answer segments for elections in defined markets to show high authority information related to the election for users, localized by language and member state. Bing Search directs users to official sources for elections, such as https://elections.europa.eu/, and other high authority sources. Bing also offers an authoritative "How to Vote" answer to government election resources for elections in defined markets.

Safety Features – Visual Search: 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, Bing blurs faces before processing data contained in files provided by the user.

User Privacy Choices – GenAI: Bing empowers users to make informed decisions about their privacy choices and how their data is collected and used in GenAI experiences. Through the Microsoft Privacy

Dashboard, authenticated users can view, export, and delete their stored conversation history, including any queries or prompts submitted in Copilot Search.

These controls complement existing privacy settings, such as cookie banners and consent mechanisms, which continue to apply to generative features. Bing also applies an internal review process to examine how choices are presented, ensuring users clearly understand its options and the implications of its decisions. By giving users meaningful control over its data, Bing supports autonomy and accountability in the use of AI-driven technologies.

Digital Literacy: To this end, Microsoft promotes digital literacy across its services, including Bing. This work is essential in helping users critically assess online content, especially in the face of misinformation or potentially deceptive material, and complements Bing's other proactive safety efforts.

In Product Indicators – Child Safety: 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 Search, there is no summarized response to CSAM related queries

Public Service Announcements: For certain high-risk query categories, Bing displays *PSAs* at the top of search results. PSAs highlight authoritative information including reporting and support options, and/or offer clear warnings about potential risks. There are, for example, PSAs in the EU related to the illegal nature of child sexual abuse material and resources and support for queries related to suicide & self-harm. PSAs are one of Bing's most visible tools for helping users navigate sensitive or dangerous topics with clarity and care.

These messages are localized, translated, and adapted to serve users in specific markets where Bing operates. Whether it's a PSA or a warning label, the goal is the same: to provide clear, immediate context when users encounter results that may carry risk.

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.

Microsoft Advertising has terms of use and content policies are published online and easily accessible for advertisers and Bing users. These terms are updated as needed and provided to users in relevant languages.

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 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.

Microsoft Advertising 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 information manipulation. In preparation for global elections, Microsoft Research recently conducted internal research concerning information integrity and elections in the age of GenAI.

Responsible Product/Service Design

Microsoft's RAI Standard: All Microsoft products that incorporate artificial intelligence, including Copilot Search and Bing Image and Video Creator, are governed by Microsoft's RAI Standard. This standard requires comprehensive reviews and mitigation planning before AI-powered features are launched.

These reviews focus on identifying potential risks, assessing the likelihood and severity of harm, and ensuring mitigations are robust, context-appropriate, and in line with Microsoft's AI Principles. For GenAI experiences, this often includes added layers of technical safeguards, transparency mechanisms, and human oversight. By rigorously evaluating generative features before deployment, Bing prioritizes user safety and dignity, while still enabling innovation. This process reflects Bing's commitment to maintaining a respectful, secure digital environment in the face of emerging AI capabilities.

Conversational Drift: To address issues related to conversational drift, Bing has limited the number of turns per chat session in Copilot Search. This limitation helps prevent responses that are repetitive, unhelpful, or inconsistent with the intended tone. The team continues to evaluate additional approaches to mitigate this issue.

Prompt Enrichment: In cases where a user's prompt is ambiguous, Copilot Search and Bing Image and Video Creator may use the LLM to build out more details in the prompt to help ensure users receive the response they are seeking. This prompt enrichment does not rely on any knowledge of the user or their prior searches but instead on the AI model. These revised queries are visible in the user's chat history and can be deleted using in-product controls.

Targeted limitations on GenAI text and image responses: To address risks associated with Bing generating harmful content, Bing has placed specific, targeted limitations on GenAI text and image responses for certain queries. For example, when Bing classifiers identify potentially harmful content in user prompts for Copilot Search and Bing Image and Video Creator, Bing may not return generated content to users, may divert users to different topics, or redirect users to traditional search results. These measures help prevent the dissemination of harmful or inappropriate content.

User-Centered Design: User-centered design and user experiences are essential aspects 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 Search and Bing Image and Video Creator for the first time, these services offer various touchpoints to help them understand the capabilities of the system, disclose that they are powered by AI, and communicate limitations. Elements of the experience also help users better understand its interactions, including chat suggestions specific to RAI, explanations of limitations, ways to learn more about how the system works, and easily navigable references to show users the results and pages in which responses are grounded.

In-Product Indicators and Disclosures: Bing's safety philosophy includes keeping users informed when safety-related actions are taken. This includes in-product indicators and disclosures such as warnings,

disclosure footers, or icons that explain when content has been removed or flagged, and why. For example, Bing may include a notification in search results when content has been removed due to legal or policy requirements. In AI-powered experiences, Bing includes clear disclaimers that the user is interacting with an AI system, along with reminders to verify content using authoritative sources. These disclosures are written in plain language and designed to be understandable even by younger or less technical users.

AI Transparency: Transparency is essential in building trust in AI. For AI Transparency, Bing displays prominent disclosures and in-product notices when users are engaging with AI-generated content. For example, in Copilot Search, an information icon near the feedback buttons alerts users that the response was generated by AI. Hovering over this icon reveals a plain-language explanation, with links to more detailed documentation. These disclosures are written at a fifth-grade reading level to ensure they are accessible to young users and audiences of varying digital fluency. To support informed engagement and reduce manipulation, Bing has also introduced content integrity as a Service, a system that uses digital credentials and cryptography to authenticate the origin of media. This helps users trust that the images or media they are seeing have not been tampered with, and that they come from verified sources.

Ongoing Monitoring of Risks: 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, One Compliance ("1CS") assessments, Security Assessment, and Privacy reviews. In addition to risk assessments, when any features are being created, metrics are evaluated and updated regularly.

Authoritative Content: Bing actively directs users to trusted sources through several tools. For high-risk events such as various EU elections, Bing highlights answers and PSAs that point to expert-vetted information. During events like Public Health crises, Bing would apply defensive search interventions such as demoting low authority results in ranking on queries that may lead users to data voids and misleading content informed by threat intelligence.

Safety Measures – News: 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 content that meets Bing News Publisher Guidelines. Bing balances this prioritization by working with in-market teams to identify high trust news sources for content that is returned in results for users queries and attaching Content Credentials to help users identify the reliability of the sources they are accessing.

Periodic Defining of New Metric Targets: Bing 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 they consider 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 for especially key areas of content concern. Finally, Bing sends externally facing daily updates on key areas of content and conduct concern to ensure appropriate prioritization particularly for improvements to GenAI features.

Red Team Testing: Bing also requires extensive red team testing to explore safety risks under pressure. These tests simulate how systems might be misused or manipulated in the real world. Multidisciplinary teams and at times, invited external experts, rigorously test how Bing Search, Copilot Search, and Bing Image and Video Creator respond to adversarial inputs or edge cases.

Red team findings are used to refine training data, harden safeguards, and improve system behavior. Bing

ensures that these tests continue even after launch, maintaining vigilance as systems evolve.

Shiproom review: Every change to Bing, whether in core search or newer features like Copilot Search, must also pass a "shiproom" review. This structured launch-readiness evaluation ensures proposed updates meet rigorous standards across safety, privacy, security, and accessibility.

Shiproom reviews require product teams to demonstrate the impact of proposed changes on Bing's core safety metrics, such as DDR. Teams must provide evidence of both offline and live (A/B) testing and explain how they've addressed any regressions or new risks uncovered during testing. Further, in the monthly shiprooms, issues and risks are presented to Bing leadership, enabling leadership ongoing governance and oversight. These checkpoints ensure that safety remains a central consideration in how Bing evolves.

Risk Management

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

Targeted Interventions

Content and Activity Moderation

Safety Measures – Defensive Search Interventions: As part of the Safety Measures – Defensive Search Interventions mitigation, Bing has a dedicated team accountable for implementing algorithmic defensive search interventions (as well as metrics monitoring and remediation) to address high impact issues in search results, such as information manipulation, hate speech, and other problematic content that could negatively impact user safety.

Bing uses query classifiers to identify queries that may be associated with elevated risk, such as those that could surface degrading, harmful, or exploitative content, and those trigger a set of defensive search interventions. These algorithmic measures are integrated rather than standalone and work together to uphold Bing's relevance, quality, and credibility standards—particularly in cases where harmful content may exploit data voids or manipulate user intent.

When a query is flagged as potentially risky, Bing may respond with measures such as boosting high-authority sources or applying algorithmic dampening to deprioritize low-quality or harmful content. These interventions are applied in accordance with Bing's core search principles and are actively monitored by the dedicated team that manages implementation and assesses impact across high-risk areas. This combined approach allows Bing to detect potentially harmful search patterns early and respond with targeted mitigations that preserve both user safety and access to high-quality information.

Safeguards – Shopping: Shopping uses a two-filtering system where Ads team conduct a first vetting and Bing conducts a second vetting process (Content 95% comes from Ads and 5% comes from Shopping crawl data).

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.

Data Governance

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. 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.

External Engagement

Redirecting Search Queries – Copilot Search: For Copilot Search, queries related to self-harm or suicide are considered high-risk, and Copilot Search 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 Search, there is a high authority approach taken regarding queries relating to eating disorders, to explain the negative effects of certain associated behaviors.

External Partnerships related to Public Security: Bing takes a multi-stakeholder approach by having External Partnerships related to Public Security. Additionally, as a member of the Global Internet Forum on Counterterrorism (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 and works closely with the ISD to understand risks related to 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.

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.

Partnership on AI: Microsoft, including key members of the Bing Search team, actively participates in the Partnership on AI ("PAI"), collaborating with peers across industry and civil society to identify and mitigate potential harms associated with synthetic media. As part of this work, Microsoft has contributed to the drafting of the PAI's Synthetic Media Code of Conduct; which are guidelines that promote the ethical and responsible creation, distribution, and use of AI-generated content, including deepfakes and digitally altered media.

Microsoft Threat Intelligence for Nation State Information Operations: Microsoft works to identify and track nation-state information operations targeting democracies across the world and works with trusted third-party partners, to provide early indicators of potential threats that can be leveraged to inform early detection and defensive search strategies.

Governance

Threat Intelligence: Bing has a Threat Intelligence team that specializes in identifying and responding to high-risk search topics. This team continuously investigates where and how search results might be falling short and applies corrections with precision. The team toolkit includes boosting the visibility of authoritative websites, demoting those with low reliability, and restricting suggestions that could lead users toward harmful or misleading content.

Policies & Other Mechanisms to Counter Harms

Directing to Authoritative Sources – Search: In cases where Bing detects threats to search quality in fast-moving or sensitive areas, Bing will direct users to authoritative sources. This includes areas such as elections, public health, or active manipulation of the information environment. Bing may demote low-authority content, boost more credible sources, restrict problematic autocomplete suggestions, or make manual adjustments to search results to limit harm. Each step is narrowly applied, with the goal of improving user safety while preserving open access to information.

User education – GenAI search results: Bing offers tools to help users understand the context and trustworthiness of search results. Bing works to be transparent about how its safety systems work for GenAI features, and for novel systems such as Copilot Search 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.

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 Search 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.

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.

Responsible Product/Service Design

Limit Lengths for Generative Search Queries: Copilot Search is limited by the 2000 character query length limit.

Bing's Safety Playbook: Bing has a defined set of review protocols, including policies for handling reported violations and escalations in Bing's Safety Playbook. Bing's Content Moderation team adheres to the guidelines in the playbook in handling violations and escalations.

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.

Risk Management

Safety Measures – GenAI – Elections: Novel elections classifier intended to restrict Copilot Search 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.

AI Generated Content Detection: Microsoft is harnessing the data science and technical capabilities of its philanthropic research center focused on social benefits of AI as well as Microsoft's threat intelligence for nation state information operations to better detect abusive synthetic media and other harmful content on the Internet. Furthermore, Microsoft's digital crimes function is investing in new threat intelligence work to pursue the early detection of AI-powered criminal activity.

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

Content Moderation – Webmaster: 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.

Monitoring Emerging 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.

Content Moderation – Advertising – Continuous 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.

Content Moderation

Compliance Management

Content Moderation – Training and Wellness – Child Safety: Bing's Content Moderation Ops team carries out training for content reviewers 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.

Content and Activity Moderation

Content Moderation – Search Content Removal: Bing maintains rigorous and principled processes, policies, and procedures for content removal and maintains policies and other mechanisms to counter harms. Central to these processes is a firm commitment to the prevention of CSEAI. Bing proactively scans content to detect known CSEAI using hash-matching technologies such as PhotoDNA and MD5. This scanning occurs before content enters the search index and is used to ensure such material does not appear in results at all, across Core Search and Copilot Search. Bing also proactively scans upload image uploads in

Bing Image Creator and Bing Video Creator to prevent known CSEAI from being uploaded. This proactive detection is supplemented by intelligence from trusted partners, including NGOs and law enforcement, and through user reporting channels. When CSEAI is identified, Bing removes it globally and without hesitation. This is a zero-tolerance area where Microsoft's commitment is absolute.

Bing also works to remove adult or sexually explicit images of another person online when shared without that person's consent. This is commonly referred to as nonconsensual intimate imagery (NCII), or "revenge porn". Microsoft considers this to be a gross invasion of personal privacy – whether the content is real or a "deepfake" image that is created using AI or other photo editing tools

Finally, Bing also removes certain other content under global legal standards or Microsoft policies, such as materials containing extremely sensitive personal information that could be exploited for fraud or identity theft. Removals may be prompted by legal orders from governments, user-generated reports, or policy-based decisions following internal investigations. Microsoft applies rigorous standards to each request, including the potential impact on user rights and public interest.

Report a Concern: Bing provides users with an accessible report a concern mechanism on every page of Bing Search and across GenAI experiences. Through this function, users can raise concerns about potentially unlawful or harmful content. These reports are routed to human moderation teams who assess submissions based on Microsoft's policies, relevant laws, and the specific context of the reported material. While not all reports lead to content removal—especially in cases involving lawful but offensive or disputed content—each report contributes to Bing's broader understanding of emerging risks and informs future updates to its safety interventions.

Account Suspension – Advertising: 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 its policies and reserves the right to remove those that are non-compliant.

Content Moderation – News – Proactive Detection: Bing scans News content using standard Bing classifiers to flag content for action, and to filter out lower quality sources. There is a continuous process of optimizing classifiers for precision, recall, coverage, and reducing defects without any significant fundamental ranking architecture changes. 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. 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.

Social Listening for GenAI Features: Bing maintains a social listening pipeline where insights and user feedback on Bing's GenAI features are collected from the open Internet, from users and potential users. 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, to ensure the insights feed into product evaluation and further improvements. These reports inform Bing as to the public perception of mitigations and serve as a barometer on topics concerning Bing's operations.

- Microsoft proactively tracks tactics and methods that users employ to jailbreak Copilot Search and Bing Image and Video Creator to produce content with a negative impact on user safety and takes daily action to address any gaps in safeguards.

- Microsoft regularly reviews the prevalence of issues within Bing Image and Video Creator 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.

Content Moderation – Tooling: 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.

External Engagement

Content Moderation – Content Removal Policies and Procedures: Bing will remove materials identified as illegal based on practice scanning, in limited cases, and following investigations upon notice.

Bing proactively scans content for CSEAI using PhotoDNA. Bing proactively scans content for CSEAI using hash-matching technologies (including PhotoDNA and MD5) to detect matches to known CSEAI, to avoid it from appearing in the search index. Bing also relies on reactive reporting as well as threat intelligence provided by third-party expert partners to identify and remove CSEAI from the services.

Bing also takes steps to globally remove (based on Bing policies) certain other 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. Bing may remove content based on 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, GenAI Experiences, and Ancillary Search Features.

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.

Right to be Forgotten: Bing's commitment to privacy and dignity is also reflected in its implementation of the Right to Be Forgotten ("RTBF"). Under this policy, users in jurisdictions such as the EU can request the removal of search results associated with their name if those results are outdated, inaccurate, irrelevant, or excessive. Requests are reviewed by trained staff using clear criteria to balance the individual's right to privacy with the public's right to access information. Where appropriate, Bing removes URLs and associated search suggestions from the index in response to validated requests. Escalation pathways are in place for complex cases, including collaboration with local courts and data protection authorities where necessary.

This process is underpinned by strong training protocols and quality assurance systems. Bing ensures consistency and fairness in RTBF decisions, while maintaining transparency about the process through user guidance and regular engagement with regulatory authorities.

Content Moderation – Travel Stories: Travel Stories are provided and moderated through a third-party partnership. This party maintains content moderation and quality assurance processes that uphold Bing's standards. Licensors commit 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.

Content Removal – Legal Reasons: 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 regarding 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.

Responsible Product/Service Design

Notice to Users: Bing's approach for Notice to Users 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 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 GenAI features, users receive in-product notice informing them of access restrictions or suspensions.

Content Moderation – Advertising – Proactive 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:
    - Advertiser Identity Verification (AIV): designed to enhance user trust and safety by requiring advertisers to prove their legal identity before they can run ads. This program acts as a deterrent to bad actors who might attempt to damage the network through various malicious activities such as phishing scams, personal information fraud, and advertising disallowed products.
    - 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 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 changes 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: Adult, Gambling, Tobacco, Weapons, Disallowed Supplements and Health Products, and Sensitive Topics (hate speech, violence, suffering, gore.)
- 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.

Content Moderation – Shopping: 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.

Risk Management

Content Moderation – Advertising: Understanding the critical importance of maintaining a trustworthy ad ecosystem, 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 processes using algorithmic systems and a robust filtration system to detect and mitigate harmful online behavior such as fraud, phishing, malware, or compromised accounts. Bing users can also report advertising for takedown through the "feedback" function. Detection of 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.

Report a Concern – Advertisements: 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.

Incident Response

Content and Activity Moderation

Safety Measures – Family Safety Settings: Parents can use the Microsoft Family Safety Center to allow parents to track their family's search history and monitor screen time. Child users whose parents have access to their search history are aware when these controls are in place.

Incident Response and Service Level Agreements (SLA): At the core of Incident Response and Service Level Agreements (SLAs) is a well-established infrastructure for responding to notices involving illegal or policy-violating content. This includes materials such as CSEAI, NCII, exposed personal information, and other high-risk violations. When Bing receives a credible report from governments, users, or trusted third parties, it acts swiftly to evaluate and address the concern.

Bing's support teams are trained in legal and policy standards, with clear procedures for escalating complex cases to local legal experts and subject matter specialists. Reports are tracked through Microsoft's internal incident management system, which ensures accountability and drives collaboration across engineering, legal, safety, and policy teams. Alerts are sent to designated stakeholders, and all incidents are triaged according to priority, with time-sensitive issues handled under strict SLAs that form part of Bing's broader Objectives and Key Results (OKR).

The prioritization process considers the nature of the report, the sensitivity of the policy area, and contextual details such as region, language, and media format. While response times vary based on these factors, most reports involving high-priority concerns are resolved within 24 hours. This fast-paced yet thoughtful process ensures that Bing's response mechanisms remain both effective and user-respecting.

Regular testing and quality checks are in place to ensure that these workflows operate reliably. Bing's commitment to continuous improvement means that incident response systems are refined over time based on lessons learned, stakeholder feedback, and evolving safety challenges.

Suspension of User Accounts: In addition to its core search product, Bing also supports services such as advertising that involve direct user participation. In these contexts, Bing enforces clear standards to prevent misuse that could harm users or compromise the integrity of the service. Advertisers who engage in fraudulent, misleading, or illegal behavior are subject to immediate suspension of advertiser accounts.

Bing applies these actions judiciously and in alignment with Microsoft's broader safety commitments. Suspension decisions are made based on evidence and internal enforcement guidelines, with appropriate documentation and oversight to ensure consistency and due diligence.

Appeals: Because Bing is a search engine, not a content-hosting service, its users do not post content directly to the service. As a result, Bing does not maintain user-facing appeals processes for content removals. However, where Bing applies enforcement actions, such as when removing or downranking search results in accordance with policy or legal obligations, it offers webmasters a means of appeal.

Through Bing Webmaster Tools, site owners who believe their content has been incorrectly impacted can request a review. These appeals are evaluated by trained reviewers, supported by escalation paths to experts where needed, to ensure outcomes are consistent with Bing's principles and commitments to fairness, transparency, and trust.

Account Restriction and Appeal Process – Bing Image and Video Creator: In serious cases, Bing may restrict a user's access to generative services or suspend a user's account when there is evidence of repeated or intentional misuse. This includes attempts to bypass content safety protections or efforts to generate highly sensitive material. User's who demonstrate clear intent to create CSEAI will have their accounts suspended.demonstrate clear intent to create CSEAI will have their accounts suspended.

For Bing Image and Video Creator, users whose prompts are repeatedly flagged for policy violations may be automatically suspended for a defined period or permanently banned in severe cases. To support fairness and transparency, these users are given an opportunity to appeal such decisions.

Appeals are reviewed by trained human moderators who consider the user's behavior, applicable terms, and context. Based on the outcome, the enforcement decision may be upheld, or the user's access reinstated. This process ensures that even enforcement actions within GenAI experiences are guided by a commitment to procedural fairness and proportionality.

Research and Partnership

Content and Activity Moderation

Partnerships to Identify Potential Threats: Bing partners with trusted independent third parties, such as Non-Governmental Organizations and threat intelligence providers, to inform defensive search mitigations.

Collaboration with third parties and adherence to external evaluations further reinforce Bing's commitment to its 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.

External Engagement

Microsoft-wide Election Support: An important component of the dedicated cross-functional team focused democracy and elections 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 dedicated cross-functional teams focused on democracy and elections are not specific to Bing (and are part of a whole-of-company approach), they can support and inform risk mitigation on the Bing service. Recent measures 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.
- 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 Coalition for C2PA digital credentials. Though this service is Microsoft wide, it helps reduce content provenance with political candidates.

- Deceptive AI Election Reporting tool. During the Assessment 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.
- Cyber-security and cyber-influence:
    - Cyber-influence 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.
    - Security Guidance: Microsoft works closely with its elections-related customers to provide security guidance and tools to improve its cyber-resilience and protect the integrity of the electoral process.
    - International Foundation for Electoral Systems ("IFES") and NDI Cybersecurity: Microsoft is supporting IFES to strengthen the cybersecurity practices of investigative journalists reporting abuse of state resources in elections. Microsoft is also partnering with the NDI.
    - NDI to strengthen the cybersecurity support infrastructure for political parties and campaigns internationally

Researcher & Academic Support: 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.
- Other Resources for Researchers: Microsoft provides a variety of datasets and tools for researchers, including the MS MARCO and ORCAS datasets, and promotes RAI practices with tools like the mitigation's library.

Microsoft Research: 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, RAI, and information manipulation. Bing is evaluating additional research

data sharing opportunities pertaining to elections.

External Engagement: Both Bing and specialized cross-Microsoft teams regularly engage with external stakeholders in areas of key policy priorities, such as violative content, information integrity, RAI 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.

External Engagement – Child Safety: 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, 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, data of minors' 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 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.

Regulator Engagement: Bing regularly meets with regulators around the world to understand key concerns, share information, and incorporate feedback into product design and safety systems as appropriate.

Societal Resilience Fund: 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.

Researcher Program to enable EU Researchers: Bing operates a 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 consistent with the recent Delegated Regulation.

Governance

Dedicated Cross-Functional Teams of Experts: Bing collaborates closely with dedicated cross-functional teams of experts across Microsoft. These partners specialize in areas such as digital safety, election integrity, and RAI. They help develop and maintain company-wide policies that guide Bing's decision-making on safety interventions and mitigation strategies. Their input ensures that Bing is aligned with evolving regulatory expectations and public interest standards.

Dedicated Expert Teams – Hate: Bing employs a specific team that is dedicated to deploying targeted algorithmic interventions to reduce the prevalence of hate speech, harassment against protected groups, and other problematic content in search.

Transparency and Reporting

Data Governance

Data Retention and Deletion Policies – Copilot Search: Copilot Search has data retention and deletion policies to help ensure that personal data collected through these features is kept as long as needed.

Transparency Reports: 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.

User Transparency: 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.

External Engagement

EU Code of Conduct on Disinformation ("COCD"), formerly known as the Code of Practice on Disinformation: Bing is a signatory to the EU COCD, formerly known as the Code of Practice on Disinformation, and subscribes to many commitments in the code dedicated to reducing disinformation risks, including providing users with indicators of content provenance, supporting good faith research into disinformation by enabling researcher access to data, and providing biannual reporting pursuant to the code on systems and policies in place to prevent disinformation on Bing across the EU, among others. Bing also provides enhanced reporting on agreed upon elections and other declared "crisis" under COPD.

The Tech Accord to Combat Deceptive Use of AI in 2024: 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 its 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 GenAI teams have worked to implement Tech Accord pillars into product safety systems.

Third Party Partnerships: Furthermore, Microsoft has developed tools and multistakeholder partnerships to combat the rise of misinformation that can pose risks to public health or negatively impact physical or mental wellbeing. Partnering with independent third-party organizations empowers 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. The coalition has an open-source content provenance tool, which allows creators to claim authorship while empowering consumers to make informed decisions about what digital media to trust.

As one example, Microsoft worked with independent third-party partners to perform red team testing of Bing Image Creator to understand the risk of misleading images generated by Bing Image Creator. As part of their analysis, the independent third party prompted Bing Image Creator to create visuals that reinforced or portrayed prominent false narratives related to politics, international affairs, and elections. Based on these evaluations, Microsoft improved existing Bing Image Creator mitigations.

Election Preparation 2024: 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.

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 Bing Ad Library: 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.

Policies & Other Mechanisms to Counter Harms

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.

Responsible Product/Service Design

Ranking Digital Rights Audit: 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.

Post-Election Reviews: Bing undertakes post-election reviews, as appropriate, to evaluate product and mitigation performance, reflect 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 CnaM to share practices & reflections.

Metrics x Funnel

Governance

Risk Governance – Roles and Responsibilities: Bing governs and manages safety risks as an essential business function. Bing follows the "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's Corporate Vice President and compliance function, annually and as part of ongoing tracking of business objectives, review how Bing is assessing and managing illegal content and other risks.
- 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 content moderation operations function has regular syncs with Bing's digital safety function, which is accountable for policy creation and evaluation, to ensure consistency in the policy understanding and enforcement (e.g., covering work items, edge cases and policy questions, emerging trends)
- Bing's compliance function is responsible for privacy, security and accessibility reviews and building related policies. This function works with respective centralized cross-functional Microsoft teams (privacy, security, and accessibility) responsible for the creation of company-wide policies and operations.
- Bing's customer services and support function is accountable for reviewing and triaging user reports.
- Bing's digital safety function is accountable for search principles creation, evaluation, and operations.
- Bing's legal function provides support to ensure search principles are aligned with legal requirements and central Microsoft policies and standards.
- 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 and gaps.

Risk Management

Internal Metrics: Bing routinely reviews the effectiveness of Bing's safety systems through internal metrics as well as social listening channels. 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.

Bing uses the 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 GenAI features across sets of sample queries.

- The metrics are reviewed monthly in shiproom meetings with Bing leaderships 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.
- In the areas where Bing observes regressions, Bing conducts issue spotting exercises and further invests in improvements to address the mitigation gaps.

Election Monitoring and Metrics: Bing has improved its monitoring capabilities setting up specific measurement on election-related query sets:

- Bing measures the percentage of responses that are grounded in web results following Bing's authority-driven ranking in Copilot Search. In sample-based testing, Bing has found that 90%+ of Copilot Search responses that were evaluated on a set of elections queries were grounded in web results following Bing's authority-driven ranking.
- Bing monitors jailbreak attempts of Copilot Search.
- Bing also uses social listening to evaluate online conversation around potential responsible AI risks regarding election and political topics and enable mitigation actions.

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).

- 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 global elections in multiple languages as part of red teaming and applied fixes as needed.

Appendix III: Terms and Acronyms

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.
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.
CnaM	Comisiún na Meán, the media regulator in Ireland, designated Digital Services Coordinator for Ireland under the Digital Services Act.
CSAM	Child Sexual and Abuse Material, refers to all any content that contains or promotes child abuse, or sexual abuse.
CSEAI	Child Sexual Exploitation and Abuse Imagery, refers to visual media that contains or promotes child sexual exploitation or abuse.
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
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.
EU	European Union, political and economic union of 27 members states located primarily in Europe.
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.
GPAHE	Global Project Against Hate and Extremism, a nonprofit specializing in tracking and countering hate and extremism globally
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.
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.
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.
MHRA	Medicines and Healthcare products Regulatory Agency, the UK's authority on medicine and device safety.
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."
NDI	National Democratic Institute, an organization aimed at strengthening the cybersecurity support infrastructure for political parties and campaigns internationally.
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 or sometimes also called "Public Safety 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.
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.
StopNCII	Stop Non-Consensual Intimate Image Sharing, a global initiative led by the UK-based Revenge Porn Helpline.
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.
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.

Appendix IV: Summary of External Consultations and Partner Insights

The United Nations International Narcotics Control Board ("INCB") puts efforts to mitigate risks under the DSA related to Illegal Content and Activities and Mental and Physical Wellbeing. Bing engaged with the UN INCB, a treaty-mandated body under the international drug control law. The INCB provided intelligence on drug sales typologies, including key terminology, emojis, and references to dangerous substances and precursor chemicals. These insights were instrumental in improving detection mechanisms for illicit drug-related content and were incorporated into Bing's internal content moderation and alert systems to strengthen service defenses. Further, the insights were added into Bing's internal threat intelligence system (known as the DIRT store), where they are used to support detection, moderation, and safety interventions across the service.

UK Communications Regulator("OFCOM"), to address risks related to Human Dignity, hosted a regulatory workshop where Bing participated in a session focused on online safety for women and girls; with emphasis on combating harassment, intimate image abuse, and gender-based violence. Insights shared during the workshop informed Bing's understanding of vulnerability and risk in gendered online spaces and contributed to internal safety-by-design discussions aimed at better protecting marginalized user groups.

The Lucy Faithfull Foundation, UK (LFF) supports efforts to reduce Illegal Content and Activities. Bing collaborated with the LFF on data exchange related to its CSEAI PSA in the UK. The engagement aimed to assess and improve current user interactions with the PSA while exploring further enhancements, including research on deterrence strategies. This partnership provided valuable input to refine Bing's defensive measures targeting child exploitation content online.

The Medicines and Healthcare products Regulatory Agency, UK ("MHRA") helps to reduce systemic risks associated with health misinformation and illicit pharmaceutical sales. Bing consulted with the MHRA, the UK's authority on medicine and device safety to discuss respective MHRA guidance. Based on this guidance, Bing outlined plans for additional search PSAs, keyword warnings, and other interventions. Some of these actions have already been activated for high-risk search terms, while others remain on the product roadmap pending further coordination. This initiative directly supports compliance with the DSA's mandates on Mental and Physical Wellbeing.

Bing coordinated with Stop Non-Consensual Intimate Image Sharing ("StopNCII") to bolster user protections related to Private and Family Life, particularly in cases involving the non-consensual sharing of intimate images. The DTSP discussion on DSA Reports fulfilled parts of an industry-wide effort to operationalize DSA requirements as Bing took part in a DTSP convening focused on DSA reporting and risk assessment practices. The forum gathered stakeholders from civil society, academia, and major tech companies to exchange methodologies and insights. Participation contributed to the refinement of Bing's systemic risk assessment approach and supported alignment with regulatory expectations on transparency and accountability, addressing all risk areas.

Dr. Dan Reidenberg (Suicide Prevention Expert, U.S.) was consulted in December 2024 on how to improve service responses to users in crisis, particularly regarding self-harm and suicide ("SSI"). Bing engaged with Dr. Dan Reidenberg, a leading expert in suicide prevention to gain further insight and recommendations, of which Bing implemented revised safety banner messaging across several English-speaking markets. These changes enhanced the effectiveness and empathy of Bing's PSAs, ensuring that individuals searching for

sensitive content are met with supportive, clinically informed resources. This initiative falls under the DSA risk area of Mental and Physical Wellbeing.

Bing continued its collaboration with Thorn, an organization committed to defending children from sexual abuse through technology. In alignment with DSA mandates on illegal content, Bing requested updated access to Thorn's CSAM keyword datasets to inform Bing's PSA triggers and moderation lists to remain current and effective. Such engagement can strengthen Bing's ability to detect and disrupt harmful content involving child exploitation across its search and media services.

To further its DSA compliance framework, Bing participated in a joint workshop hosted by the Global Network Initiative ("GNI") and DTSP; which convened industry, civil society, and academic experts to discuss systemic risk assessment under the DSA. This collaboration facilitated a multi-stakeholder exchange of best practices and reinforced Bing's risk identification and evaluation processes across multiple service domains.

To address election-related systemic risks under the DSA, Bing initiated a partnership with the DPA, Germany's national press agency, to receive timely and verified fact-checking data in the run-up to the 2025 elections. This collaboration was designed to inform defensive search interventions targeting information manipulation and manipulative narratives. Implementation began in April 2025, marking a significant step in Bing's civic integrity efforts within the EU regulatory context.

Bing contributed to the Safer Internet Webinar (Child & Youth Safety in AI), which focused on the implications of artificial intelligence for child and youth safety. The session explored how AI technologies shape online experiences for minors, identifying emerging risks such as harmful content amplification, exploitation, and inappropriate personalization. Takeaways from the session informed Bing's trust and safety roadmap, particularly regarding protective interventions in child-oriented search and content environments.

Bing has continued to engage with CnaM through regular meetings throughout the year. During the last year, Bing met with CnaM regarding Article 40's requirements to enable researcher access to data and explained the challenges associated with data minimization and the complexity of data requests in the search context. Bing also responded to a voluntary RFI from CnaM concerning its efforts to support media literacy in Ireland ahead of Ireland's 2024 national elections.

The European Commission held a workshop concerning the DSA systemic risk assessment process and publications requirements for both assessment and audit reports during October 2024 with designated services. The Commission reaffirmed that its only guidance will be in the form of investigations and RFIs. The Commission also noted to the services that redactions in publication must be limited and legally defensible.