

Bing Systemic Risk Assessment Report

August 2026



Contents

Contents 2
Executive summary 4
- Bing's principles and risk considerations 4
- Managing user safety in a dynamic information environment 4
- Continuous improvement and assessment results 5

Introduction 6
- Purpose 6
- Document structure 7

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 22
- How Bing monitors effectiveness of risk mitigation efforts 26

Bing Overview and Risk Profile 27
- User Base and Market 27
- Considerations for Bing's Risk Profile 27
- Core Search Features 29
- Generative AI experiences 34
- Ancillary search features risk profile 42
- Changes to Bing's features 44
- Systemic Risk Assessment Methodology Overview 46
- Risk assessment process 47
- Assessment tooling 50

Risk assessment results 50
- Overall results 50
- Changes to Bing's risk profile 52
- Emerging trends and shifts in global risk posture 53
- Changes in risk rating 54
- Looking ahead to the coming year 112

Appendix I: Detailed systemic risk assessment Methodology 114

Risk areas....114
Risk factors and influencers....115
Risk assessment inputs....115
Probability....117
Severity....118
Inherent risk....119
Maturity of mitigations....120
Residual risk....122
Sample scoring calculation....123
5C Typology....123
Appendix II: Bing's digital safety risk and compliance framework and mitigations....124
Product quality....124
Product safety....128
Targeted interventions....138
Content moderation....144
Risk Management....149
Incident Response....150
Research and Partnership....152
Transparency and Reporting....156
Metrics x Funnel....158
Appendix III: Terms and Acronyms....161
Appendix IV: Summary of External Consultations and Partner Insights....164

Executive summary

Bing's principles and risk considerations

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 among 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. Bing must balance these interests and rights to maximize benefits to users while minimizing possible harms.

This approach is based on Bing's Trustworthy Search Principles and a supporting framework designed for balancing these competing values. The Trustworthy Search Principles include:

- 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.

Because search engines play a critical role in enabling access to information, Bing seeks to balance the mitigation of potential harm with the protection of fundamental rights. Over-moderation of content in search could have a significant negative impact on the right to access information, freedom of expression, and media pluralism.

Bing increasingly leverages AI to help people find information faster and more intuitively — including through more conversational ways of asking and receiving results. As users increasingly engage with AI-enabled search experiences, Bing adapts to help them discover, evaluate, and access relevant information.

As these interactions evolve, Bing remains at its core a search service: a gateway for finding and accessing information hosted by third parties. Unlike digital services primarily designed to facilitate open-ended conversations, social interaction, content sharing, gaming, or sustained online engagement, Bing's primary function is to help people discover and access relevant information in response to users' queries.

This distinction shapes how Bing addresses content risks. Amidst rapid changes in user behaviors, search patterns, the adoption of generative AI, and new formats and capabilities, Bing remains committed to its mission and principles to help people access relevant information while mitigating potential harms. Addressing content risks in search requires a different approach than may be appropriate for other types of online services that facilitate sustained online engagement.

Managing user safety in a dynamic information environment

Bing uses a robust, multi-layered system of mitigations to deliver this commitment to manage user safety. This approach is organized around five core functions and driven by metrics, striving to

continuously improve. The mitigations 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 people's feedback and reports to provide a safe experience for people who use Bing services. Teams across Bing continuously measure the effectiveness of applied risk mitigations in alignment with these principles to adjust and refine them as threats and risks evolve.

Consistent with our commitment to ongoing improvement, Bing made progress in each of the five enhanced mitigation areas identified in the 2025 Systemic Risk Assessment (SRA). These include strengthening Bing's capabilities to detect and manage dynamic and complex risks, enhancing preparedness and response capabilities for rapidly evolving events, and improving protections for children through SafeSearch. For example, Bing improved the audit traceability of product changes that may affect Bing's Digital Service Act (DSA) systemic risks, developed more detailed policy implementation guidance to enable safe generative experiences where possible, strengthened its operational capabilities to monitor, detect, and respond to rapidly changing threats and measure ongoing effectiveness for nuanced harms.

Importantly, however, Bing as a service does not exist in a vacuum but within a rapidly evolving ecosystem shaped by technological advances, geopolitical developments, regulatory priorities, and changing societal expectations. As part of its continuous monitoring activities, Bing tracks these developments to identify emerging risks and adapt to mitigations accordingly.

Concerns relating to Child Sexual Exploitation and Abuse Imagery (CSEAI), Non-Consensual Intimate Imagery (NCII), fraud, and other forms of harmful content continue to evolve as increasingly capable GenAI tools enable malicious actors to operate with greater speed, scale, and sophistication. At the same time, expectations for online platforms have expanded, particularly around strengthening protections for minors and preventing access to explicit content. Bing also monitors evolving views on digital services and features that encourage sustained online engagement, which may introduce new risk considerations.

To better understand emerging harms and assess mitigation effectiveness, Bing engages with researchers, academics, and civil society organizations, collectively referred to as partners. Bing incorporates these insights into its ongoing risk management approach. For example, in this assessment cycle, Bing's collaborations with partners resulted in greater localization safety interventions, including public service announcements addressing suicide and self-harm, eating disorders, NCII, and child safety.

Bing uses information on trends from its continuous monitoring activities, including engagement with external partners and stakeholders, 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 ecosystem developments, stakeholder insights and Bing usage trends is necessary and key to keeping Bing services safe for the people who use them. Annual assessments are also utilized to reflect 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 DSA.

Continuous improvement and assessment results

The results of Bing's continuous monitoring and annual SRA, factoring in external shifts and internal product changes, indicate that residual risks to people, stemming from use, misuse, or functioning of Bing services, remain consistent and primarily fall within the Low range. Bing assesses that the

mitigation measures in place across all risk areas are proportionate, reasonable, and effective in reducing risks on the service to an acceptable level, based on implementation of available scientific and technical insights and the incorporation of internal and external data and expert opinion.

In the 2026 SRA, Bing assessed eight of the twelve 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. 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 remains committed to continually enhancing people's 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.

As it has in each risk assessment year, as part of this 2026 SRA, Bing identified the following five opportunities to enhance users' trust and safety:

1. Launch continuous monitoring dashboard to bring together pre- and post-launch metrics and enable regulatory tracking and reporting;
2. Continue advancing threat intelligence and interventions, in particular related to NCII and Fraud, to respond to changes in the ecosystem;
3. Continue to engage with the broader ecosystem to deepen the understanding of unique safety risk profile and challenges for search products and features;
4. Update Bing's risk assessment workbook that supports our SRA, including by streamlining mitigations and the mitigation maturity assessment, to enable greater clarity on Bing's layered system and highlight what is changing over time; and
5. Better align regulatory cadences with core business processes to evolve from a once-annual process – with intermittent Critical Impact Risk Assessments (CIRAs) as required – to regularized, routine, recognized, and rewarded risk assessment and audit processes.

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 (SRA), 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 Bing Overview section. This includes core search features; ancillary services, including search verticals such as Images, Shopping, Maps, News, Travel, and Advertisements in Bing; and GenAI-enhanced experiences, including 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 2025 - July 2026.

The Report discusses mitigations implemented to manage 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 thorough overview of Bing's approach to protecting users online, 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.

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 (2) its approach to protecting fundamental rights and (3) the multi-layered system of mitigations Bing uses to manage user safety.
- **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 SRA including an overview of the methodology used, a description of the risk assessment process, and an explanation of the assessment tools, resources, and inputs 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 prior 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 SRA, 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

Search engines play a vital role in society, enabling people to access information and conduct research, and upholding the fundamental rights to Freedom of Expression and Information and Media Pluralism. Consistent with not unduly restricting users' ability to search and receive information, Bing balances these rights against other vital interests – privacy, safety, democracy, public health, and national security – as reflected in the Trustworthy Search Principles below.

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 help confirm 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 to 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 support narrowly tailored actions, 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 actions taken 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?

These principles also underpin Bing's GenAI experiences, which represent a very small portion of users' experiences. Bing continues to evolve its policies, operations, and infrastructure to better identify, assess, and manage the risks associated with GenAI. These updates continue to operate within Bing's Trustworthy Search Principles and the multi-layered system of mitigations Bing uses to balance access to information with user safety.

Additional information on Bing's Trustworthy Search Principles can be found at How Bing Delivers Search Results and in Bing's Webmaster Guidelines. Additional information on Bing's generative AI experiences is available in the Microsoft Services Agreement.

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 in 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 uses a robust and multi-layered system of mitigations to manage risks and maintain this careful balance. These mitigations 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
5. Incident Response

Each layer of the Funnel includes 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 measures the presence of content that violates Bing policies and the efficacy of implemented safety mitigations. DDR allows Bing to evaluate the effectiveness of its mitigations, identify regressions, and take corrective action. In addition to overall DDR, Bing 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. The sections then 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.

Advances in user safety in the last year

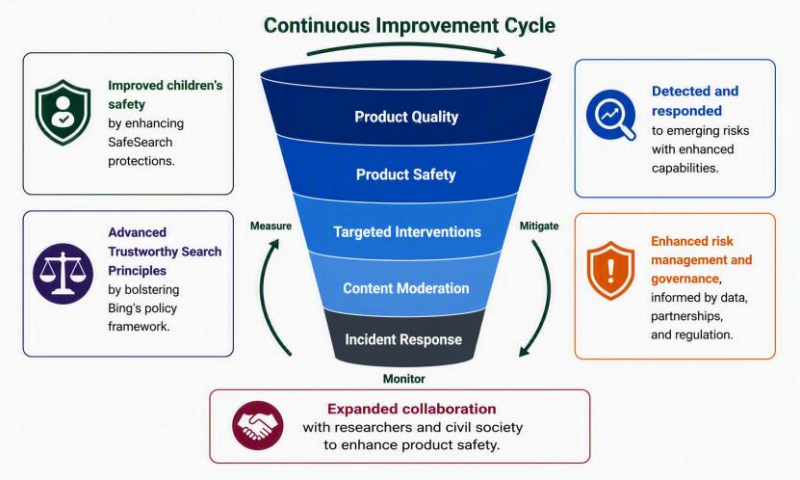
Bing's safety mitigations are organized around core functions and are continuously measured, monitored, and refined through a culture of continuous improvement. As users increasingly engage with AI-enabled search experiences, Bing's focus remains on sustaining and strengthening the effectiveness of its safeguards in a manner aligned with the role of search as a gateway, rather than a destination for open-ended social or conversational interactions.

Over the reporting period, and consistent with the culture of continuous improvement, Bing continued to strengthen its ability to identify, assess, and manage systemic risks, while preserving access to lawful information. Bing is highlighting significant advancements made this year that are consistent with the enhanced mitigation commitments described in Bing's 2025 Systemic Risk Assessment report. These include improving how Bing:

- Protects minors online, using SafeSearch modes; described in the ***SafeSearch Mitigation***;
- Implements the Trustworthy Search Principles using a multi-dimensional policy framework to balance access and risk; described in the ***Policy Mitigation***;
- Detects and manages dynamic, complex risks, including by
    - Enabling Bing to prepare for and respond to events, using Events Playbook; a new mitigation that complements the existing ***Threat Intelligence Mitigation***;
    - Adapting to risks related to so-called nudification apps; and
    - Supporting vulnerable individuals in need by increasing the scope, coverage, and localization of PSAs; and
- Assesses and manages risks, based on data, partnerships, and regulation, in alignment with DSA
    - Incorporating the Article 28 5C Topology;
    - Enhancing product governance; and
    - Tracking operational data (e.g., action rate, Turn Around Time (TAT)) for incident response and Bing's limited content moderation
- Partners with researchers and civil society to improve risk management

Each of these advances will be discussed within the layer of the Bing Safety Funnel below.

Figure 2: Enhanced mitigations across layers of the Bing Safety Funnel



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 enable people to receive high-quality results that align with their search intent while limiting exposure to harmful or low-quality content.

A 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 recommend content based on user behavior across the service, Bing's approach centers on providing relevant results based on a user's query.

When Bing surfaces suggestions, such as autocomplete, trending searches, or homepage modules, these are carefully filtered through targeted mitigations and quality criteria to promote relevance, authority, and safety. This supports 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, relevant, fresh, and 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, outdated or manipulated content. Rather than removing content outright, Bing relies ***on continuous improvement of ranking algorithms*** to help prevent harmful content from being surfaced prominently, particularly when users have not explicitly searched for it.

To support this, Bing continuously invests 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 people, Microsoft's policy teams, regulators, and civil society organizations to support its ranking decisions to reflect evolving standards of trust and accuracy.

Bing's ranking function continues to support user protection and fundamental rights by prioritizing relevance, freshness, and authority in search results. During the assessment period, Bing continued

to improve ranking quality through the use of semantic ranking technologies, including both small and large language models, in support of the same core ranking purpose: identifying relevant indexed web content and determining the order in which results and partner-generated answer types are presented to people. These ranking improvements are intended to better understand user intent, surface current information, and elevate trusted, high-authority sources where appropriate, while preserving users' ability to access a broad range of lawful information.

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. That is, Bing relies on ranking algorithms to prevent harmful content from being surfaced prominently, particularly when users have not explicitly searched for it, rather than removing content outright.

To support this, Bing uses several factors, including Quality and Credibility (QC) scores. QC score 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.

Advanced Trustworthy Principles in Practice: Example Safety Guardrails Based on Risk and Intent

Bing applies safety guardrails based on both the risk associated with a category of content and the degree to which user intent can be reasonably inferred from the query. In practice, this means Bing does not treat all potentially sensitive queries the same. Instead, Bing evaluates a range of signals, including query specificity, commonly observed query patterns, intent indicators, and quality and safety considerations, to determine the appropriate search experience.

For example, Bing considers adult-related queries searching for legal pornographic content (e.g., "porn vid" or "porn pic") to be High Risk. When such user queries are vague and broad (e.g. searches involving anatomical terms such as"breasts" or "penis") or not specific enough, Bing prioritizes high-authority content and does not predominantly rank explicit content high in the search engine results page. This is because Bing cannot confidently determine from the query alone that the user is specifically looking for pornographic or other explicit content. In such cases, Bing leans on providing results that are high-authority and not likely to result in harm (e.g., definition of "breasts" or "penis", news articles, etc.). SafeSearch setting-based and jurisdiction-specific filtering may also apply.

By contrast, when user queries are specific (e.g., "porn xx website" or "naked woman pic"), Bing respects user intent to find such content/information and ranks results accordingly, while continuing to apply Bing's quality, credibility, and safety signals to support responsible and lawful information access. Access to such content may nevertheless be limited for certain users, including through Strict SafeSearch and other family safety settings, age-appropriate experiences, unauthenticated experiences, or where required by applicable law.

These ranking systems are continuously measured and refined to help align with 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. Users can also adjust advertising and content personalization settings. For example, if an individual frequently searches for travel content, Bing might personalize their Microsoft News feed with travel-related stories. Microsoft provides users with controls over certain personalized experiences and data uses, including advertising and content personalization settings where available. In the EU, Bing also requires users to opt in before using cookies or similar technologies for advertising personalization and other purposes that require consent under applicable law.

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

Enhancing Protection for Minors: Default SafeSearch for Under-18 Users

In Europe, Bing defaults known users under the age of 18 to SafeSearch Strict Mode.

This approach leverages an existing safety feature to reduce exposure to explicit content and reflects Bing's current, practical application of age-informed protections, while broader age assurance capabilities continue to evolve.

SafeSearch settings can be locked to prevent tampering and enable consistent application of filters. Bing also provides a mechanism for users to report inappropriate content or suggestions, including explicit content, NCII-related concerns, and other sensitive material. These reports feed into the continuous improvement of the underlying safety systems, helping Bing respond quickly to emerging issues and regulatory expectations.

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, the mitigation helps individuals 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 Bing can respond quickly and effectively to emerging safety challenges across the service.

Engineering processes and safety by design

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 enable 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 verify 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 supports integrating safety, privacy, and integrity 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, summarized search features, and Bing Image and Video Creator respond to adversarial inputs or edge cases.

Red teaming and systematic sweeps are conducted to identify and address defects, particularly in advance of high-risk situations, such as elections or major breaking news events. For example, during the initial weeks of the Iran war, Bing conducted targeted reviews as part of Bing's Defensive Playbook for elections and crisis events. Bing also performs systematic sweeps focusing on emerging problem areas, such as nudification tooling in search. Red team findings are used to refine training data, harden safeguards, and improve system behavior. Bing confirms 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 assesses whether proposed updates meet rigorous standards across safety, privacy, security, and accessibility.

Balancing Tradeoffs Between Safety and Access to Information

Bing applies a multi-dimensional framework that considers the topic, the user's apparent intent, and the feature involved to determine proportionate mitigations. Rather than broadly removing content, Bing may apply targeted measures such as ranking adjustments, SafeSearch filtering, content warnings, promotion of authoritative sources, or modified generative responses, rather than broadly removing content. See "Balancing Tradeoffs Between Safety and Access to Information" earlier in this report.

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 with ongoing governance and oversight. These checkpoints help support safety as a central consideration in how Bing evolves.

Product policy and safety features

As a search engine, Bing is committed to supporting and ***prioritizing users' fundamental rights to search*** and their rights to access and seek information. Bing does not broadly remove content from its index. Instead, Bing focuses on minimizing unintended exposure through ranking and targeted filtering. Bing enhanced the policy framework that guides how the service balances user safety, legal obligations, content quality and relevance, and access to information when determining how content is surfaced in search experiences. Consistent with this approach to delivering search results, Bing crawls, indexes, and ranks content to provide users with relevant, credible, and authoritative information responsive to their queries, while applying targeted interventions where appropriate to mitigate potential harms. This approach enables Bing to support open access to information while taking proportionate steps to address identified risks. Additional information on Bing's approach is available in How Bing delivers search results.

Bing recognizes that safety includes not only what individuals 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 people in making informed decisions.

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

Expanded and Localized PSAs Through Partnerships

During the assessment period, Bing expanded and localized Public Safety Announcements (PSAs) to provide users with access to authoritative resources related to ***Eating Disorders, Suicide, Self-Harm, and NCII.*** *These efforts help connect people to appropriate health information, crisis-support services, reporting mechanisms, victim-support resources, and other trusted assistance that supports users' mental and physical wellbeing.*

Bing works with local and regional partners, including civil society organizations, trust and safety experts, and victim-support groups, to help ensure these resources are relevant and effective across markets. Through these collaborations, Bing localizes PSA content to reflect local languages, available support services, legal frameworks, and cultural considerations, helping ensure users receive information and assistance that is appropriate to their circumstances.

For certain high-risk query categories, PSAs are one of Bing's most visible tools for helping users navigate sensitive or dangerous topics with clarity and care. 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. Representatives from product, policy, safety, and partnership teams collaborate to define applicable query sets, identify suitable resources, and determine the markets in which the intervention is appropriate. These efforts are often supported and informed by civil society and interested local partners. PSA messages and destination resources

are then localized, translated, and adapted to the legal, cultural, and service context of applicable countries before being deployed at the top of relevant search results. There are, for example, PSAs in the EU related to the illegal nature of Child Sexual Abuse Material (CSAM) and resources and support for queries related to suicide & self-harm. Whether it's a PSA or a warning label, the goal is the same: to provide clear, immediate context when people encounter results that may carry risk.

Bing's safety philosophy includes keeping people 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 individual 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 person.

Targeted Interventions

The third layer of Bing's Safety Funnel is Targeted Interventions. Targeted Interventions are a combination of algorithmic and manual interventions used by Bing when search results may not align with their 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 people while preserving access to information.

As a search engine, Bing intervenes in limited ways, consistent with Bing's Trustworthy Search Principles, to help individuals 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 people's safety.

In addition to threat intelligence, Bing uses classifiers such as those that 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's safety and access to high-quality information.

In cases where Bing detects threats to search quality in fast-moving or sensitive areas, Bing applies its established ranking, quality, and safety systems to elevate authoritative information and reduce

the visibility of lower-quality content. In certain circumstances, however, the novelty, speed, or potential impact of an event may require additional intervention by Bing teams to address heightened risks and help shape users' first impressions of emerging information.

In these situations, Bing may direct users to authoritative sources, such as on select high risk elections. Bing may also demote low-authority content, restrict problematic autocomplete suggestions or summarizations, or make manual adjustments to search for results to limit harm. These interventions supplement Bing's standard approach and are applied when necessary to address specific risks that may not be fully captured through automated systems alone.

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 through proprietary detection pipelines and processes, as well as paid partnerships with factchecking and threat research organizations. Bing continued to invest into further scaling of threat intel capabilities through expanded partnerships and ability to identify viral narratives targeted by information manipulation and detection of crisis events that carry Search risks**.** The team toolkit includes boosting the visibility of authoritative websites, demoting those with low reliability in ranking and grounding of summarizations, and restricting suggestions that could lead individuals toward harmful or misleading content.

Bing expanded its suite of threat intelligence partners to improve coverage and support broader factchecking and information integrity efforts. These partnerships provide additional signals that help Bing identify emerging risks, strengthen its response to information manipulation, and support the fact-checking ecosystem. Bing continues to use these threat intelligence inputs alongside internal red teaming, user reports, social listening, and operational monitoring to detect gaps and refine targeted interventions over time.

Expanded Framework for Emerging Crisis Events

The Crisis Playbook may be activated for significant events that create elevated information-quality or safety risks, such as elections, terrorist incidents, natural disasters, public health emergencies, armed conflicts, or rapidly emerging online events. The playbook provides a structured framework for assessing event-specific risks and determining whether additional monitoring, threat-intelligence analysis, or targeted mitigations are warranted.

Bing maintains and continues to refine its dedicated ***Crisis Playbook*** that enables rapid identification, escalation, and response to emerging real-world events that may increase search-related risks. The playbook provides structured procedures for monitoring developing situations, coordinating cross-functional response efforts, deploying targeted defects monitoring, threat intelligence and search interventions, and continuously reassessing risk as events evolve to help make sure users are presented with authoritative and reliable information during periods of heightened uncertainty.

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 measurement of search defects, 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 people's feedback on Bing's features are collected from the Internet. Flagged viral potential issues 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 a barometer of public sentiment on various topics related to Bing. In the reporting period, Bing has continued investing in optimizing the processes for alerting and responding to potential product incidents based on social reports.

Bing's Targeted Interventions help mitigate the risk of people unexpectedly 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 people's 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 authority of the requesting party.

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 enforcement of copyright takedown requests and other legal content obligations that are subject to established review processes and applicable legal requirements.

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 to prevent such content from existing in Bing's index or from appearing in search results. This scanning occurs before content enters the search index and is used to ensure such material does not appear in search results at all. 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 third-party expert 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 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. Bing continues to mitigate NCII and related harms through policy enforcement, detection, and user controls. Because distinguishing NCII from explicit content can be difficult in practice, particularly because a search engine may lack context available to other platforms to establish consent, Bing maintains strong default protections for minors and is reviewing safeguards for NCII pathways, including reporting and removal of workflows, to maintain clarity, prominence, and effectiveness across jurisdictions.

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 individual rights and public interest.

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 people with options for tailoring their content.

Bing provides people with an accessible ***report a concern*** mechanism on every page of Bing Search and across GenAI experiences. Through this function, people 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, trained staff review requests from individuals in the EU who seek removal of search results associated with their name when results are outdated, inaccurate, irrelevant, or excessive. 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 maintains 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; helping users 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 supports 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 assessed and actioned on within 24 hours. This fast-paced yet thoughtful process helps Bing's response mechanisms remain both effective and user-respecting.

Regular testing and quality checks are in place to verify 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. When Bing takes enforcement actions, such as removing or downranking search results in accordance with its policies or legal obligations, Bing provides webmasters with a mechanism to request review or appeal those actions. Bing takes this approach because it indexes and ranks content from across the web instead of hosting user-generated content directly. As a result, Bing does not maintain a separate user-facing appeals process for content removals.

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 help promote outcomes that 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 maintain 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 **Microsoft's *dedicated cross-functional teams of experts***. These partners specialize in areas such as digital safety and RAI. They help develop and maintain company-wide policies that guide Bing's decision-making on safety interventions and mitigation strategies. Their input helps ensure 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 third-party research organizations to contribute to novel research and internal studies concerning safe design practices, responsible AI, and information manipulation.

Both Bing and Microsoft's dedicated cross-functional teams of experts regularly ***engage externally with stakeholders*** in areas of key policy priorities, such as improving protections for minors,

mitigating exposure to violative content, information integrity, responsible AI. External engagement provides Bing opportunities to hear feedback from a range of civil society, non-profit, researchers, and government stakeholders and learn from others in the industry. and provide a greater understanding of concerns about Bing and broader societal and technology issues. As part of external engagement, Bing meets with regulators, election officials, and other public authorities to understand election-related risks. Engagements inform Bing's internal policies, practices, and standards. These engagements also help Bing identify, assess, and manage risks.

Expanded Partnership Engagement and External Collaboration

During the reporting period, Bing continued to engage with a broad range of external partners, including researchers, civil society organizations, industry groups, and independent experts, to inform risk identification and mitigation efforts. These engagements spanned partners across EU member states, including national regulatory authorities and Digital Services Coordinators including the Czech Republic, Cyprus, Germany, Hungary, Ireland, Malta, Portugal, Sweden, and the Netherlands, as well as industry coalitions and civil society organizations engaged at the European level. Engagements with national authorities encompassed topics including elections integrity, non-consensual intimate image sharing (NCII), consumer protection and fraud, generative AI in search, age assurance, and researcher access obligations under Article 40 of the DSA. These engagements provided additional perspectives on topics such as elections, information integrity, responsible AI, child safety, synthetic media, and online harms. Insights from these collaborations helped inform Bing's ongoing assessment of emerging risks and the effectiveness of existing mitigation measures.

Bing operates a qualified ***researcher program to enable EU researchers*** to easily request access for publicly accessible Bing datasets under Article 40.4 and 40.12 of the Digital Services Act (DSA). Researchers can request access to Bing data that is publicly accessible on Bing, upon application, from a singular landing page. In addition, researchers can request access to datasets that are not publicly available through the DSA Qualified Research program.

To support the threat intelligence team, Microsoft has ***third-party partnerships*** with independent organizations, researchers, and civil society to analyze potential future harms that may arise while using Bing. Partnering with independent third-party organizations empowers Bing to implement targeted interventions to promote more authoritative information.

As one example, Bing commissioned a structured expert assessment of Copilot Search to understand how users interact with the experience and what harms can arise from it. The study ran realistic and deliberately probing prompts — including extended multi-turn sessions — logging every prompt, response and source, and assessing each interaction for safety, appropriateness and benefit. These findings translate directly into safer generative results: they are driving work on contextual signal detection, stronger intent-checking on high-risk topics, session-level rather than prompt-level risk assessment, more consistent overview coverage on sensitive queries, and clearer provenance cues on AI-generated summaries — all grounded in documented evidence.

Bing also held workshops with civil society organizations, sharing our own thinking and gathering expert perspectives in return, both on specific technology-and-society issues and on how Bing approaches risk analysis more broadly.

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

Improved Monitoring and Measurement of Moderation Operations

During the reporting period, Bing expanded its ability to measure and monitor its limited content moderation operations through more comprehensive operational metrics. These metrics include measures such as action rates, review volumes, and turnaround times (TAT), which support ongoing evaluation of moderation processes and mitigation effectiveness. Enhanced operational visibility helps Bing identify trends, assess performance against established processes, and inform continuous improvements to its risk mitigation program.

Bing publishes regular ***transparency reports***. Microsoft's Reports Hub and Digital Safety Site are publicly available sources where people can access 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 jurisdiction, community and privacy, and security transparency reports that people can easily access. For example, Bing's most recent DSA Systemic Risk Assessment and Transparency report are hosted on the Reports Hub.

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 make these resources available in relevant languages and markets.

People 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 individuals 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 enhances transparency around the ad selection process by explaining the relevance of ads based on individual online behavior and interests, thereby offering people greater insight and authority over their digital environment.

Bing is a signatory to the ***EU Code of Conduct on Disinformation (COCD)*** 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, participating in several Election Rapid Response System (ERRS) efforts under the code for elections in the period 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 COCD.

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 may include the creation of deceptive synthetic media, including exploitative images of

people, the generation of discriminatory, hateful, explicit, or otherwise harmful content, the spread of misleading information, privacy harms arising from the generation or disclosure of personal information; and misuse by bad actors. At the same time, measures intended to prevent these harms may inadvertently restrict legitimate content, creative expression, or access to information. Bing addresses these considerations through a layered set of safeguards, governance processes, and technical controls, including the mitigations described below.

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 people using Bing services, such as engaging in harmful, fraudulent, or illegal activities, violating privacy, or creating inappropriate content. Individuals 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 individual 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 people to evaluate the credibility of the information presented. On identified high-risk queries, we may apply further safeguards to source selection and response framing or not show summarizations.

These mitigations are implemented to reduce risks that GenAI outputs may inadvertently surface low-authority or misleading information, reinforce biased or incomplete perspectives, or be interpreted by people as disproportionately authoritative relative to underlying source material.

Copilot Search Safety Enhancements

During this assessment period, Bing implemented infrastructure enhancements that supports more appropriate Copilot Search responses, helping reduce the likelihood that people encounter unsafe or misleading outputs. These enhancements complement the existing Safety Funnel and GenAI-specific mitigations described in this section and reflect Bing's ongoing approach to adapting its safeguards and mitigation measures as generative AI technologies and associated risks continue to evolve.

GenAI responses operate in conjunction with Bing's broader ranking and retrieval systems, and mitigations are designed to support alignment between generated outputs and underlying search quality signals to reduce the risk of inconsistent or conflicting information presentation across experiences.

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 their 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 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 people on the nature of AI-driven search experiences, including their uses, safeguards, and limitations. This approach helps mitigate the risk that people may rely on generated content without understanding its limitations, by reinforcing transparency and enabling validation against authoritative sources.

Copilot Search responses include ***citation links***, and the corresponding search results are displayed under the generative response on Bing Search Engine Response Page (SERP). Within the Copilot Search vertical, all of the traditional search results can be viewed by clicking the "all" link at the top, left of the page. This practice allows individuals 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.

Copilot Search and Bing Image and Video Creator may use an LLM to evaluate user inputs against established Responsible AI (RAI safety policies). ***LLM-based guardrails and safety mediation*** is used to evaluate user inputs against established Responsible AI (RAI) safety policies prior to downstream content generation. For example, if a user prompts for a picture of a golden retriever riding a scooter, the guardrail may fill in details about the scene and artistic type before requesting the model to generate images in order to align the response with the user's original request. This capability operates as one of multiple safety layers that work together to help identify and address harmful, inappropriate, or borderline policy violations before any core downstream generation or content synthesis takes place. These layered mitigations are intended to reduce the likelihood that generative AI experiences produce content that violates RAI policy. This process does not rely on any knowledge of the user or their prior searches.

Bing empowers people 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 pre-launch review process prior to new or changed features being deployed to examine how choices are presented, ensuring users clearly understand their options and the implications of their decisions. By giving users meaningful control over their 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 people engage with AI-generated content. For example, in Copilot Search, an information icon near the feedback buttons alerts people that the response was generated by AI. Hovering over this icon reveals a plain-language explanation, with links to more detailed documentation. Disclosures are intentionally written to make them easy to understand by people of varying digital fluency. 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 in Bing Image and Video Creator 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, in Bing Image and Video Creator. If a user 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 AI-generated content or interactions that may not be appropriate.

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 so that 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 designed to support 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.

These safeguards also aim to mitigate risks of user misuse, including attempts to generate harmful, misleading, or policy-violating content, or to use GenAI outputs in ways that may facilitate fraud, harassment, or other harmful activities.

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 summarized search features and Bing Image and Video Creator. For instance, classifiers

have been implemented to mitigate the risks of generated content related to public figures and other categories of content that may create heightened safety, trust, or policy-compliance concerns. Bing has also invested in problem-specific classifiers to power defensive search interventions such as identification of nudification tooling related queries and results.

In 2026, this approach continued to mature through a consolidated structure that connects policy development, measurement, classifier-based algorithmic interventions, and operational moderation capacity.

A key area of enhancement was Bing's response to NCII and related notification-tooling risks. Bing strengthened its NCII policy to address risks associated with NCII content, notification tools, NCII sites, and sites with exploitative removal practices. To support this policy framework, Bing developed specific query sets that allow the team to measure at a scale where defensive leakage may occur in search results. These measurement efforts support ongoing evaluation of residual risk and help identify areas where mitigation performance should continue to improve.

Content Moderation

Copilot Search uses the same content moderation processes as Core Search, and individuals 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 includes 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 certain generative services*** that requires log-in or suspend an individual'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. Individuals 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 to 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 supports procedural fairness and proportionality for enforcement actions within GenAI experiences.

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

Defect Detection Rate (DDR): A Key Measure of Mitigation Effectiveness and Harm Reduction

Bing uses DDR as a primary means to measure 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 the quantitative signal of safety performance of DDR, Bing's commitment to metrics spans its layered system of mitigations, and each has a 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 165 million average Monthly Active Users (MAU) in the EU during the period of August 2025 to July 2026. 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. Bing features, including generative AI experiences such as summarizing articles or documents, translating, writing, and rewriting in different tones or styles, are available to both authenticated and unauthenticated users. 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. Research conducted by Bing 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 all EU Member States as of July 2025.

Considerations for Bing's Risk Profile

When using a search engine, people can have many valid reasons for wanting to seek out legal content that could be seen as 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.

As noted above, Bing and search itself are evolving, including through AI-assisted experiences and richer content interactions. However, Bing's role remains fundamentally distinct from chat services, social media platforms, and gaming environments. People generally come to Bing with a specific informational need, seek relevant information, and conclude their interaction once that need is met.

Accordingly, Bing continuously evaluates systemic risks through the lens of its search-specific functionality and risk profile, while continuing to refine its mitigation approaches in response to changes in the information ecosystem, user needs, societal expectations, and regulatory developments. 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 support appropriate stewardship of that data and minimize risks of harms related to misuse of personal data.

However, Bing generally does not host user content, allow for messaging between users, or allow users to publish content on the service. As a result, Bing presents a lower vector for systemic harms associated with user-to-user communications and user-generated content than services that enable these activities. Bing also does not provide the means for users to post or share their own content directly on the service, unlike social media platforms. Consequently, Bing faces limited risk that user behavior toward other users, the sharing of user-generated content, or the enforcement of rules governing such content will give rise to significant systemic harms. For additional information on design principles that underpin Search, please see the Design Principles Shaping the Search Experience section below.

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 has established terms and conditions that apply to its users, and enforces those terms where needed – such as in the case of an individual attempting to use illegal CSEAI as a search prompt in visual search, or individuals who violate terms or attempt to jailbreak GenAI features – user behavior on the service, on its own, is not a significant vector for systemic 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 are 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 reduce risk of harm in search results related to the content they seek.

People still cannot post or share content on Bing's generative AI experiences, but Microsoft recognizes that there is an increased risk that people may generate content using these experiences 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 experiences to support appropriate use and fair enforcement when needed.

As an ad-supported service, Bing recognizes that there are risks associated with advertisements on the service and works to support alignment of advertisements appearing on the service with 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, including mitigating the risk that advertisements may serve as a pathway to harmful, low-authority, or misleading third-party content through promoted results or embedded links.

Across the Bing service, systemic risks may arise from the combined operation of ranking systems, generative AI, vertical experiences, and advertising pathways, which together influence how people are exposed to and interact with information.

Furthermore, there is an increasing focus across the ecosystem on fraud, dark patterns, child safety, and AI, with particular emphasis on the measurement and assessment of user harm. In response, Bing is strengthening its approach to measuring consumer harm associated with fraud risks across both search and advertising experiences and is working to develop a more unified view of user outcomes across product surfaces. Bing is enhancing its measurement and reporting processes to better capture user-harm signals and to provide clearer traceability between implemented mitigations and observed user outcomes.

Bing recognizes the continued evolution of search experiences and that harmful outcomes can manifest through search and related features. Bing's approach is to be proactive, experiment and learn within applicable legal frameworks, and balance fundamental-rights tradeoffs across jurisdictions and product surfaces as expectations evolve.

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. A combination of algorithms generates Bing Search results by matching the user's search query with third-party webpages in Bing's index. In some scenarios, Bing may suggest searching 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. At most Bing 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 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

The first step in building search is figuring out which pages exist on the Internet to index them, a process commonly called crawling. Bing crawls the web to build and maintain an index of pages, or URLs, that can be displayed in response to a user-initiated search or action. Bing Search results are generated through ranking systems that match the user's query with third-party webpages in Bing's index and organize results based on relevance, quality, and alignment with user intent.

Bing continues to improve its ranking systems through iterative model and architecture enhancements. In 2026, the Core Search team described the evolution of ranking technology from BERT-based approaches toward the use of small language models for ranking relevance improvements. This evolution is characterized as a gradual and continuous improvement in ranking quality rather than a fundamental change in the core search experience. Bing's ranking approach continues to use deep-learning-based methods to assess relevance and quality signals, while leveraging newer model capabilities to improve the usefulness of results returned to users.

Bing has also simplified aspects of the core ranking stack. The Core Search team noted that the ranking pipeline was reduced from nine layers to five layers over the past year, enabled by more capable models that can process more information within each layer. This consolidation is intended to improve efficiency and search quality and does not change the basic operation of core search: users submit queries, Bing matches those queries to indexed web content, and ranking systems determine the order and presentation of results.

In some scenarios, Bing may suggest search topics or query refinements based on user activity, search history, or trending topics, including through search suggestions and related search experiences. Where such features are not owned directly by the Core Search team, they should be assessed in the relevant feature or product area while remaining part of the broader Bing search experience.

Design principles shaping the search experience

Bing's search experience is guided by design principles that reflect its nature as a user-initiated, request-driven service, prioritizing responses to specific user intent. Search is optimized to deliver relevant, high-quality, and authoritative results aligned closely with user queries, reinforcing user agency, and minimizing the amplification of content beyond what users actively seek.

These principles emphasize targeted, relevance-driven responses over engagement maximization, to support that the experience is structured to efficiently satisfy discrete information needs rather than encourage sustained engagement. Features intended to make search more efficient for users – such as autocomplete and trending suggestions – are carefully scoped to remain aligned with user intent and safety expectations.

Collectively, this approach supports a utility-focused, non-addictive interaction model, where users are able to find information and complete tasks without exposure to mechanisms designed to drive prolonged or compulsive use, while maintaining access to high-quality information. Passive interaction with Bing is limited in both nature and scope. For example, a user searching for who won a match may scroll through a finite set of results pages, after which no further content is served. Results are ordered once, based on relevance and authority; no algorithm continuously re-ranks or promotes content as the user pages through search results.

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 "Adult male giraffes typically stand 16–19.7 feet (5–6 meters) tall, while females are slightly shorter at 14–16 feet (4.3–4.8 meters)'. To enable users to quickly find the information they need, Bing may also provide a sidebar pane with source information, other related topics and further giraffe-related content and information included below. Answers are generated by Bing partner teams, including answer experiences, such as Q&A, weather, sports, and other specialized answer types. Answer eligibility is determined on a query-by-query basis, taking into account whether a specialized answer type is available and capable of satisfying the user's query intent. Core search evaluates available answer types, may generate answer text that summarizes information from web content, and determines which answers are most relevant to the user's query, whether an answer should be shown, and where it should appear in relation to other search results. Core search applies ranking principles and signals, including relevance, freshness, and authority, to help surface high-quality and current information from trusted sources. Bing may also offer specialized answers, such as in relation to elections or other high interest topics. Some answer experiences may also support or be used by generative AI experiences, including by providing relevant grounding links; those generative experiences are addressed separately in the "Generative AI experiences" section.

Figure 3: Core search results screenshot

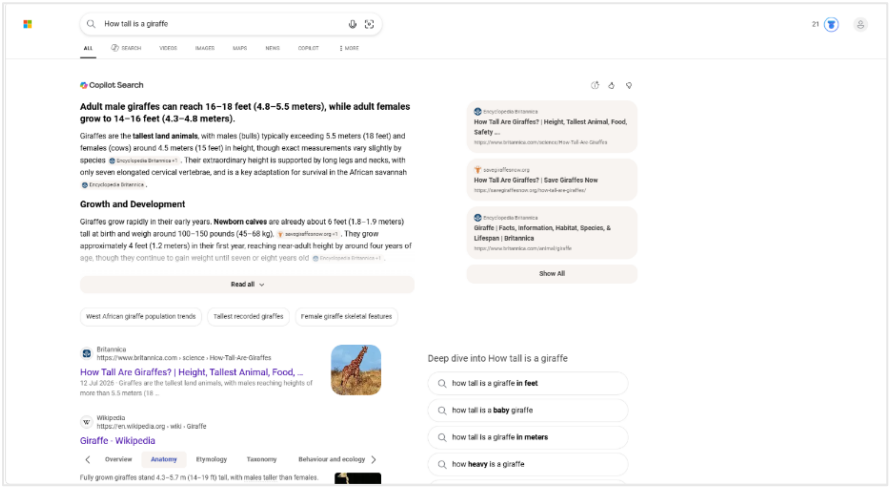
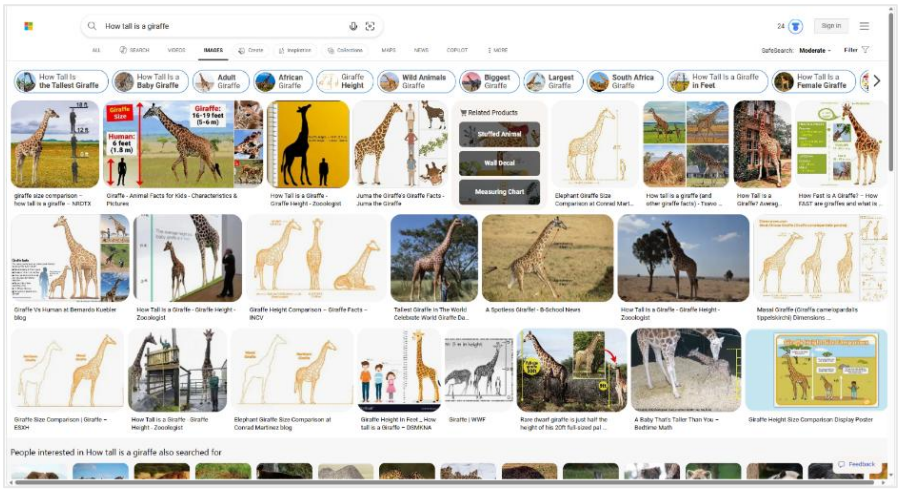


Image and video

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



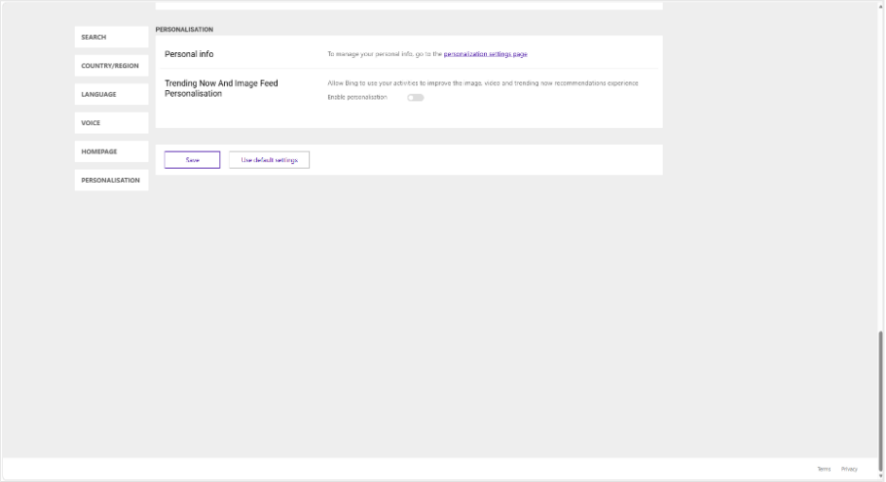
When a user first lands on the image or video experience homepages, prior to entering a query, the homepage may show recommended content based on prior interactions with the site, such as the user's search history, engagement with image results, saved images, where a user allows it, and their Edge browsing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

Bing is continuing to invest in multimodal retrieval capabilities that support a more comprehensive understanding of visual and web-based content. The Core Search team described ongoing collaboration with the image and Multimedia teams to combine core search's text-based retrieval and vector modeling with image-based retrieval capabilities. This work is intended to help Bing evaluate content more holistically by considering both textual and visual information associated with a page, rather than relying on either modality in isolation.

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 of tolerance related to safety is warranted to preserve 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. Users do not generally use Bing search results or search result URLs (such as URLs beginning with bing.com/search) as a mechanism for sharing information with others. Instead, they directly share the destination content reached through those results, such as a social media post, article, video, or shopping page they wish to share.

As a result, content is less likely to "go viral" through the Bing service itself. Therefore, while referenced harms may be present within core search, the impact is unlikely to become systemic due to the limited nature of 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.

The Core Search team's role is focused on ranking — identifying relevant indexed web content, applying ranking models and signals, and determining the order and presentation of results and

partner-generated answer types. Authority is a core ranking consideration, and core search incorporates authority-related signals as part of its broader relevance and quality evaluation when ordering results. Core search does not generate content, own the underlying index, or independently create harm-specific safety signals. Instead, it integrates signals and interventions from partner teams within Bing, including safety, defensive, and spam teams, into ranking systems.

For the 2026 assessment period, the core search risk profile remains broadly stable. The Core Search team did not identify new risk areas, material risk escalations, or known incidents specific to core search since the prior assessment period. Core search continues to present systemic risk considerations primarily through the possibility that ranking systems may surface, prioritize, or fail to sufficiently demote third-party content that is low quality, misleading, harmful, illegal, or otherwise inconsistent with Bing's safety objectives. These risks are mitigated through Bing's broader layered safety approach, including ranking quality measures, authority and freshness signals, partner-team safety signals, and service-wide interventions.

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

Risks by feature
Search experiences
• 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 searches 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, 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 above.

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, 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 OpenAI, 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 people find what they need and represents a significant advancement in transforming the search experience. Copilot Search responds to the user query with generative text, which 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 that is grounded in the retrieved search results. 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 the ranking of web sites in the search results that are used by grounding.

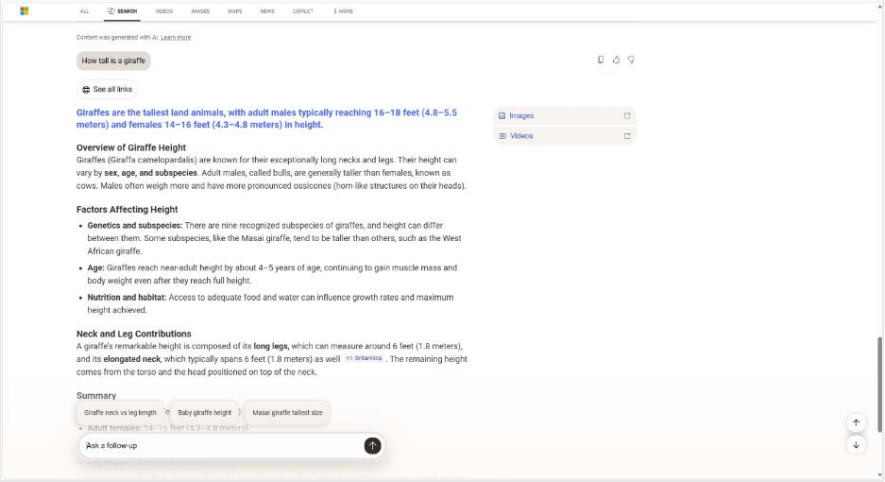
Bing has designed Copilot Search with publishers and content owners in mind to support a healthy web ecosystem. In addition to the citations within the generative responses, Bing includes inline links for the entire sentence or passage within the responses so the user can easily navigate to the sources. 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 multiturn experience. Once a user clicks on a topic, results are quickly summarized, and additional follow-up searches will be suggested so that 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. Copilot Search supports
follow-up prompts; however, the experience does not maintain persistent conversational state across turns. That is, each follow-up is treated as a new, standalone search that carries limited context from the prior query, rather than an ongoing conversation that accumulates history. This design reduces susceptibility to iterative multiturn jailbreak attempts by limiting cross-turn context accumulation within Copilot Search. Copilot Search will, in some instances, not provide a GenAI experience because responses may be inconsistent with Bing's policies and/or applicable law (for example, queries on certain sensitive topics or that seek restricted content). In these cases, Copilot Search will return traditional core search results related to the user's prompt.

Copilot Search continues to implement enhanced safeguards, including topic policies that limit GenAI responses on sensitive areas, real-time classifiers to flag and filter problematic content, and embedded Responsible AI instructions via meta prompts to decline harmful or illegal queries. As Bing deepens integration of generative features into Copilot Search, safeguards emphasize high-authority grounding and "safety generation" controls to support summaries that are high-quality, safe, and appropriately cited to authoritative sources.

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.

In the reporting period, Copilot Search added features that support 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 who are not signed in have a 30-follow-up questions/queries limit per day whereas users who are signed-in have a 100-follow-up questions/ queries 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 or video, supported by Bing Image or Video Creator. This image creation functionality is the same as what exists on bing.com. Copilot Search was also moved to a new orchestration engine that changed where responses are generated while maintaining the same safety classifiers, filters, and enforcement stance. These engineering changes did not alter the applicable policies or the user safety posture of the experience.

Figure 6: Copilot 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 7: Search suggestions screenshot (core search)

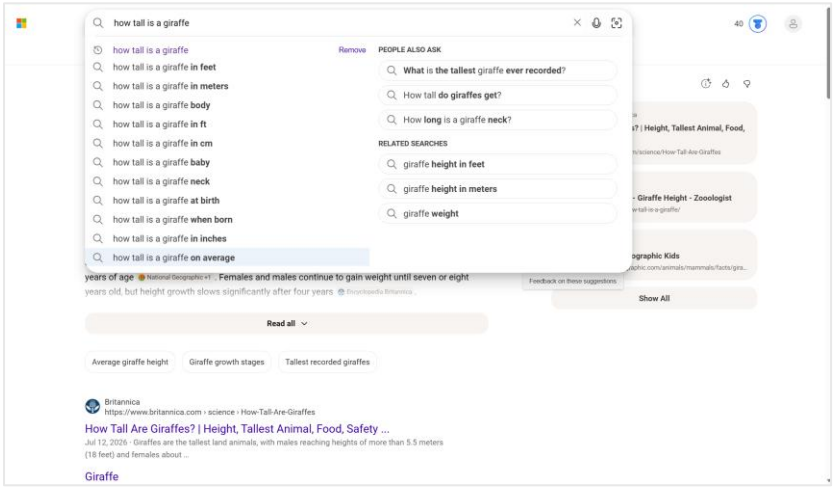
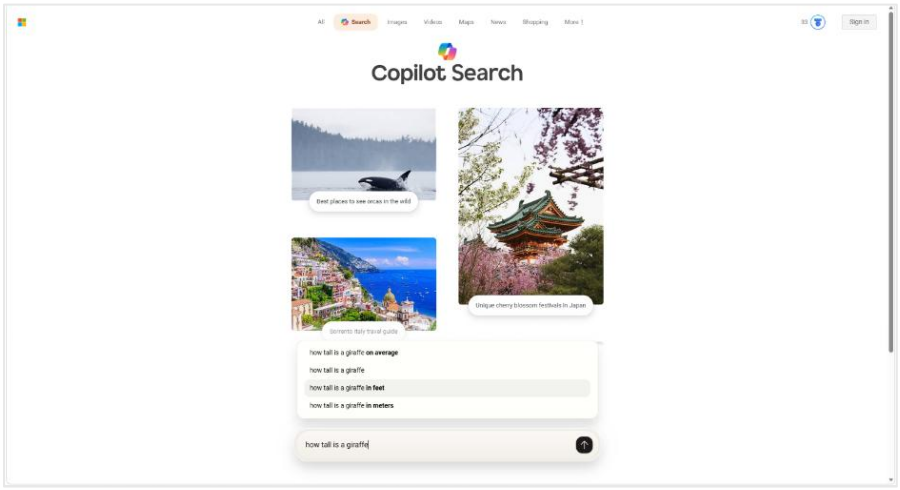


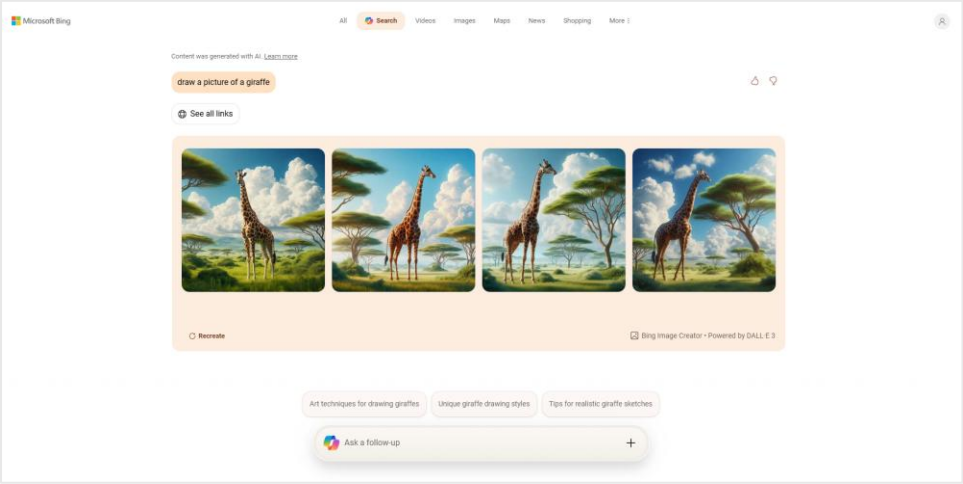
Figure 8: Search suggestions screenshot (Copilot Search)



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. During the reporting period, Bing Image Creator shifted from an image generation model from OpenAI to an image generation model from Microsoft, 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 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. Bing Image Creator is generally not available to users signed in with a work or school account.

Figure 9: 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 with audio capabilities. 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. During the reporting period, Bing Video Creator added a feature to enable users to upload images to inspire videos. This feature leverages the same safety mitigation stack used by Bing Image Creator to prevent users from uploading CSAM into the Bing index and to mitigate the risk of the model returning a video inconsistent with the users' request.

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:

Risks by feature
GenAI experiences
• 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 by 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

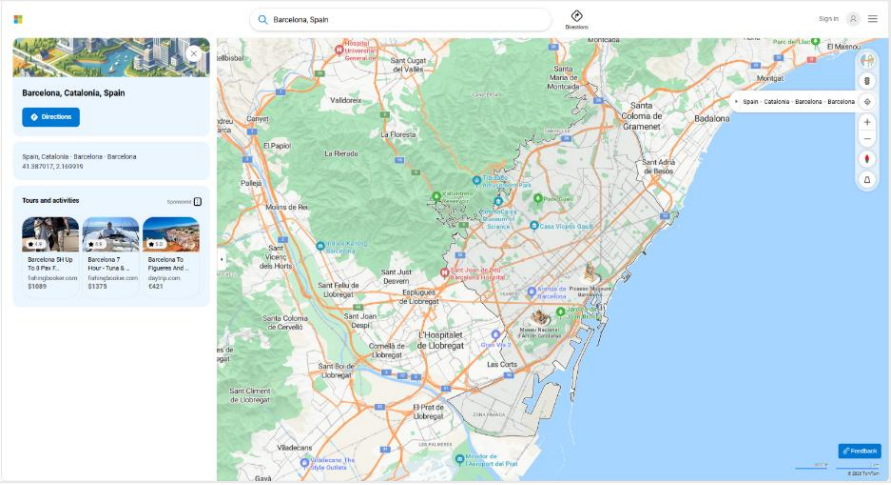
Bing Core Search links to a number of "verticals" or features areas to enable enhanced search functionality:

Maps

Bing Maps allows people to search for information about businesses, landmarks, locations, and other geographic data in a map interface. Bing Maps is an ancillary search feature available at bing.com/maps that enables users to search for locations, places, businesses, cities, and addresses, obtain directions between locations, and view map information through visualization modes, such as road, satellite, and traffic views. 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 first- and third-party apps on their mobile device. Personal data is not shared with these apps unless the user explicitly chooses to share it.

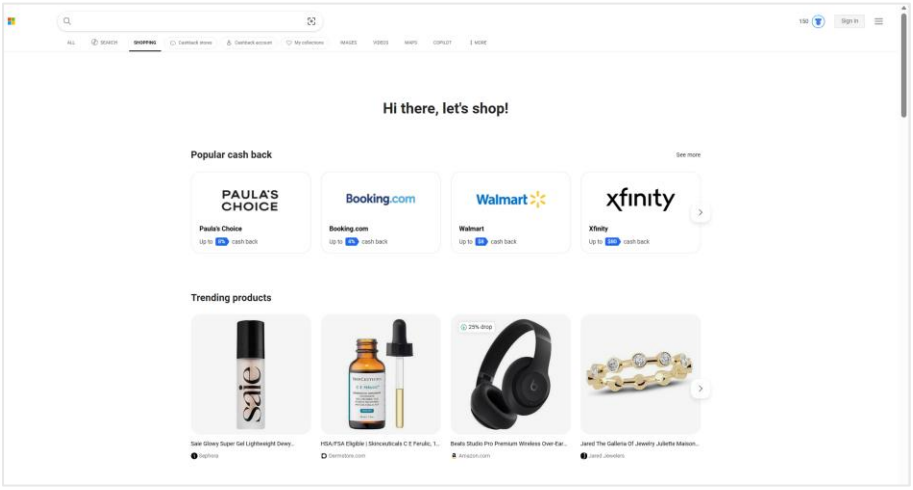
Figure 10: Bing Maps interface screenshot



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 the organic (unpaid) results shown to the user below paid advertisements; paid advertisements are clearly labeled and shown separately. 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.

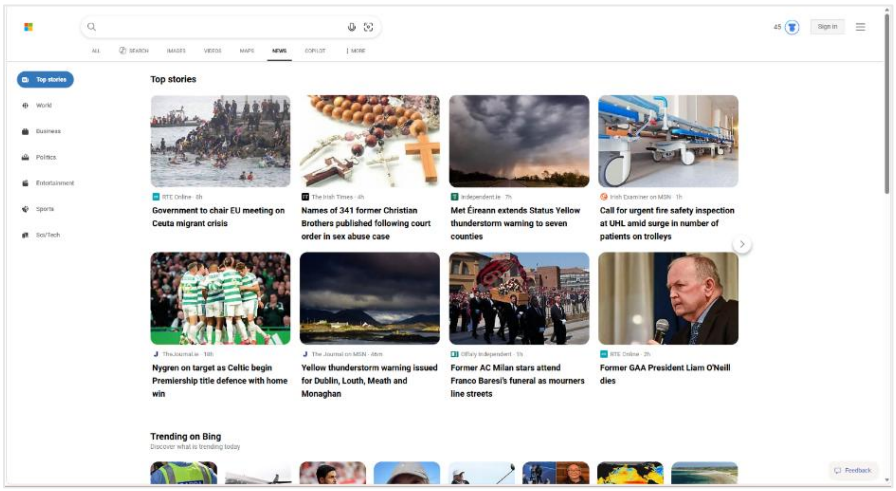
Figure 11: Bing Shopping landing page screenshot



News

Bing News is a dedicated news vertical that 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.

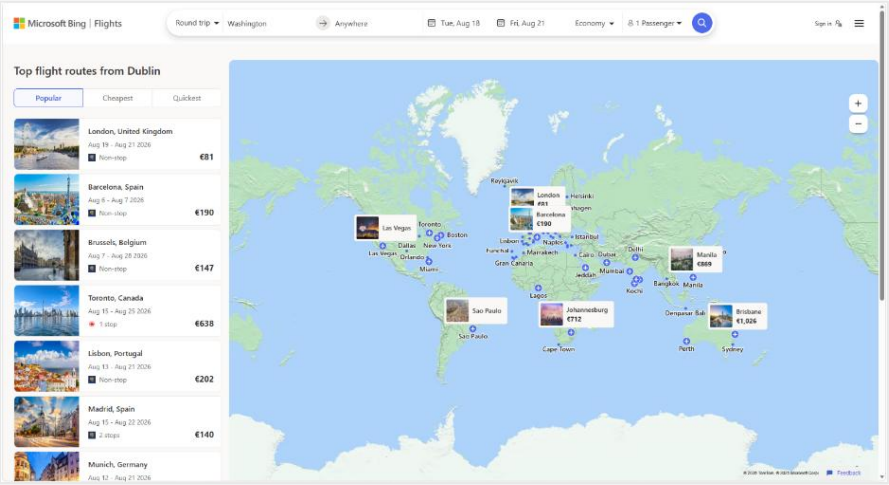
Figure 12: Bing News search vertical



Travel

Bing Travel (also currently known as "Bing Flights") is a specialized search and trip-planning tool that helps people explore destinations, plan trips, and get travel insights. Bing Travel has narrowed into a single, structured search experience. Users can search for and compare pricing for travel information, such as flights. 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 payment or booking transactions. Note, many of the Bing Travel features, including the Bing "Travel Stories" feature (introduced in 2025 as a vertical photo and video-style content under Bing Travel) are no longer part of the Bing Travel offering.

Figure 13: 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 (e.g., as "Ads" or "Sponsored") and are displayed to users primarily based on their search queries and other user data, such as browsing behaviors and inferred traits. 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:

Risks by feature
Maps
• Risk that Maps could provide incorrect information related to polling locations. • Risk that Maps shows content that perpetuates bias, reinforces stereotypes, or discriminates against products, vendors, or businesses. • Risk that Maps recommends sites or 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 that do not conceal their privacy are shown in the Travel the World Virtually experience or in Maps features.
Shopping
• Risk that Shopping over-moderates and removes available products related to political parties or figures in a manner that perpetuates political bias.

Risks by feature
• Risk that users could input targeted queries seeking products with slogans or graphics that include discriminatory content and view such products in the first few pages of Shopping 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 or include gender-based violence. • Risk that activities on Shopping are used to collect funds to finance terrorist organizations.
News
• Risk that News returns low-authority or low-quality news sources containing election-related misinformation. • Risk that third-party web content linked in News results includes hate speech, discriminatory content, harmful stereotypes, or content that facilitates discrimination toward individuals or groups. • Risk that, to comply with government orders, News may block access to third-party web content in a manner that unnecessarily censors legitimate access to information. • Risk that News results include adult or age-inappropriate content. • Risk that News shows content that instructs on, promotes, normalizes, praises, or facilitates illegal activities or materials. • Risk that News results show content that promotes, normalizes, praises, or facilitates self-harm or harm to others. • Risk that News results include misinformation related to crisis events or public security.
Travel
• Risk that Travel provides insufficient information about travel products or services, preventing informed consumer decisions. • Risk that Travel shows locations, listings, or destinations associated with adult services. • Risk that Travel ideas, itineraries, or search results suggest dangerous routes, destinations, or activities that place individuals in physical danger or undermine physical well-being. • Risk that private property is featured as a destination or activity in Travel results. • Risk that Travel offerings or coupons expose users to phishing, hacking, or data breaches. • Risk that Travel shows locations or rental listings used 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 lack transparency in political advertising targeting practices in a manner that may 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 targeting, affecting protected groups' ability to access critical services. • Risk that adult or age-inappropriate advertisements appear on Bing.

Risks by feature

- Risk that advertisements on Bing promote illegal content or activities.
- Risk that advertisements on Bing promote addictive or harmful products, including illegal substances, negatively impacting mental or physical well-being.
- Risk that advertisements on Bing promote legal but harmful substances.

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's 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's trust and safety and compliance 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 a new feature or feature update on Bing's DSA systemic risk profile.

Feature updates since the 2025 Systemic Risk Assessment report are discussed below.

Copilot Search

As noted above, Bing is continually innovating and evolving the experiences to delight its users. The evolution of Bing's generative AI offerings since the 2025 DSA Systemic Risk Assessment represents the dynamism of search and generative AI.

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. Follow up prompt handling was clarified this year – Copilot Search does not maintain persistent conversational memory across turns. Instead, follow-up prompts are handled by reformulating the user's next query rather than carrying forward a conversation history. That is, each follow-up is treated as a new, standalone search that carries limited context from the prior query, rather than an ongoing conversation that accumulates history. This design reduces susceptibility to iterative multiturn jailbreak attempts by limiting cross-turn context accumulation within Copilot Search.

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.

Copilot Search provides suggestion features that guide user refinement throughout the Copilot Search and follow-up experiences; while placement has changed since last year, this functionality continues to be surfaced to users to encourage relevant next steps.

Taken together, these changes reflect a year of incremental improvements: backend orchestration has been modernized while safety mitigations remained constant; the UX presents suggestion features with updated placement; and multiturn follow ups are explicitly constrained by design to avoid persistent conversational carryover that could otherwise be exploited across turns.

This integration of GenAI within the Bing suite of services is aligned with Bing's Trustworthy Search Principles to support 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 June 2025 with text prompt functionality and expanded in September 2025 to support image uploads as inputs. Bing Video Creator is similar to Bing Image Creator in that it generates videos based on the prompt the user inputs. 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.

The feature 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 risks such as the generation of misleading, harmful, or sensitive visual content. By expanding into video content, Bing is broadening the scope of its generative service while maintaining a focus on safe and responsible deployment. Further, as of March 2026, the Sora 2 video generation model is now powering Bing Video Creator. The Sora 2 model enables more dynamic motion generation, richer visual fidelity, more photorealistic outputs, and integrated audio capabilities, representing a material enhancement in output quality and user experience.

In connection with the introduction of audio, the team evaluated audio-specific harms and introduced policy adjustments to address them. Protections include technology that evaluates the generated audio via transcript analysis to identify policy-relevant concerns, as well as technology to recognize copyrighted music in outputs to help prevent misuse. These updates complement Bing's existing prompt and content-level protections applied to the creation of workflows.

Maps

Between 2025 and 2026, Bing Maps underwent a significant redesign that streamlined the product and focused the experience on its core uses, including location search, business discovery, and directions. As part of this effort, Bing retired or simplified a number of legacy features and lower-usage capabilities.

This included the deprecation of the Local Guide feature, which had enabled users in France and Germany to explore hotels, bars, restaurants, events, and other points of interest; this feature was officially deprecated in October 2025 and is no longer part of the Bing Maps offering. The updated Maps experience reflects a shift toward a more streamlined and focused set of functionalities, centered on core use cases, such as location search, business discovery, and directions, while reducing feature complexity and retiring lower-usage capabilities, including collections (My Places), print functionality, and certain 3D or flyover experiences. In addition, the redesign incorporates architectural improvements, including a modernized codebase, enhanced state persistence through URL sharing, and reduced interface complexity, aimed at improving usability and performance. Overall, these updates reflect Bing's broader approach to iterating product features to enhance people's experience while maintaining alignment with its core search and navigation objectives.

Travel

Over the past year, Bing Travel has undergone material deprecation and simplification as a standalone vertical, reflecting a focus on core search and AI-enhanced search experiences. Historically, Bing Travel functioned as a more feature-rich, trip-planning surface, offering aggregated travel data alongside curated elements, such as "Travel Stories", which provided destination-focused multimedia content sourced from third-party partners. However, during the current assessment period, Bing removed or scaled back several of these differentiated, content-heavy features, resulting in the discontinuation of more immersive or editorialized travel experiences.

As a result of these updates, Bing Travel no longer operates as a distinct, fully featured product vertical, but instead persists as a lightweight aggregation layer within Bing Search. Users can still access
travel-related information – such as flights, accommodations, and destination insights – but these experiences are now primarily limited to surfacing third-party content and booking options via external links, rather than providing integrated discovery or storytelling functionality within Bing itself.

Systemic Risk Assessment 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 SRA, 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 relevant risk factors, 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.

Bing leveraged its established SRA approach, with targeted improvements to enhance transparency, consistency, and interpretability. These updates also reflect a deliberate effort to streamline and simplify the overall methodology, reduce unnecessary complexity, and improve the usability of the framework for consistent application across teams and reporting periods. These updates include refinements to the assessment of likelihood (probability), emphasizing clearly defined, evidence-based inputs from internal metrics, external data sources, and expert judgment, and reducing reliance on bespoke composite measures that were less directly interpretable. Bing has also rationalized select methodological components, including consolidating overlapping mitigation categorizations and removing elements that did not materially affect risk. Collectively, these updates reflect Bing's ongoing commitment to continuous methodological improvement, helping the assessment to remain robust, proportionate, and grounded in available scientific and technical insights, while supporting clarity and consistency in the interpretation of results across reporting periods.

The methodology draws on a structured set of mitigation measures and measurement frameworks applied across Bing services. Service-wide mitigations are described in the Bing Safety Funnel and detailed in Appendix II, while the metrics and measurement framework used to evaluate mitigation effectiveness, including DDR and related indicators, are described in the "How Bing monitors

effectiveness of risk mitigation efforts" section. These components together inform the probability, severity, and mitigation of maturity assessments applied across risk areas.

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 well-being
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 multidisciplinary stakeholder input and information on updates to Bing features since July 2025, on the presentation and prevalence of risks across features, and on the implemented mitigations and ongoing monitoring activities across industry standard practices – 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.

In accordance with the latest 5C Typology issued by the European Union, the Bing Risk Assessment team supplemented its DSA-aligned framework with the 5C typology as an additional input to consider when identifying and defining risks related to minors. The 5C typology helped structure consideration of how risks to children may arise across Bing's products and services, including through exposure to harmful content, harmful contact, harmful conduct, consumer-related harms, and broader cross-cutting or compounding harms. Notably, Bing does not have features that enable harmful contact or conduct. Bing used this typology to support risk discovery, as a prompt in stakeholder inputs, and as part of assessing how minor-related risks may manifest across core search, generative AI experiences, and ancillary features. The 5C typology did not replace the report's twelve risk areas but rather informed the identification and articulation of minors-related risks within and across those existing DSA-aligned categories.

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 the use or misuse of Bing. The assessment considers both the service's inherent vulnerability in the absence of mitigation measures and user demand or likely frequency of attempts to perpetrate each harm. 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 multilayered 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 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 the relevant defensive classifiers trigger on less than 1% of overall user queries. Bing uses this trigger rate to contextualize raw DDR results observed within high-risk adversarial query sets and to better represent risk across overall search traffic.

Severity

The Bing Risk Assessment team continues to focus on objective evaluation, 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 well-being), while considering the scale (individual to global) and gravity (or remendability) of impact to arrive at a severity score and rating.

Risks with an 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 well-being) 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 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 2026 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.

Within each residual risk category, ordering reflects inherent risk characteristics and mitigation maturity, where relevant. In the 2026 DSA Systemic Risk Assessment, the Residual Risk Ratings for Bing remain low or minimal across most assessed areas, with one risk area rated minimal, eight risk areas rated low, and three risk areas rated moderate. Freedom of expression and information is rated as a minimal residual risk. The three risk areas rated as moderate residual risk are human dignity, illegal content and activities, and private and family life. These moderate-rated areas reflect risks that are evolving, higher-impact, or subject to heightened public and regulatory attention, while remaining within an acceptable range given Bing’s implemented mitigations and its unique risk profile as a search engine. Bing assesses that current mitigations are reasonable, proportionate,

and effective, and it will continue to monitor and refine mitigation strategies to maintain compliance with its systemic risk obligations under the DSA.

Bing considers the 2026 ratings to be reasonable and supportable based on the current assessment methodology, available internal and external signals, and the mitigations in place during the assessment period. Bing will continue to review these ratings as additional data, regulatory expectations, and risk signals evolve.

Figure 13: 2026 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	Likely	High	High	Optimized	Low
Civic Discourse and Electoral Processes	Highly Likely	Critical	Critical	Optimized	Low
Consumer Protection and Fraud	Likely	High	High	Managed	Low
Discrimination and Hate	Not Likely	High	Moderate	Managed	Low
Mental and Physical Wellbeing	Likely	Critical	High	Optimized	Low
Private and Family Life	Likely	High	High	Managed	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	Managed	Low

Changes to Bing's risk profile

Over the past year, Bing has continued to operate within a dynamic and evolving risk environment shaped by technological, geopolitical, and societal changes. While individual factors in the risk ratings have shifted in select areas, Bing's overall systemic risk profile remains broadly stable due to continued investment in mitigation capabilities and governance.

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. At the same time, Bing recognizes that search experiences and the broader digital ecosystem are changing, and it continues to refine its understanding of systemic risks and the measures used to address them.

This consistency within Bing's ecosystem, combined with the steady application of safety and relevance principles, contributes to a stable systemic risk profile. Building on last year's effort to monitor DDR aligned with DSA systemic risk areas, in this reporting period, Bing built out its ability to measure content moderation and incident response operations. These enhancements will enable greater real-time monitoring and 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 risks associated with image-based abuse and the nonconsensual sexual exploitation of individuals. Bing developed additional implementation guidance, metrics, and technology to better manage NCII risks. Together, these investments improve Bing's ability to detect and manage potential NCII, including by improving identification of risky queries; reducing visibility of nudification tooling and potentially harmful content in related search results; and improving management of search suggestions and generative summarizations for identified NCII-related queries to reduce the risk of directing users to NCII content, tooling, or instructions. As noted above in How Bing manages User Safety, Bing has begun implementing PSAs in the search results page in certain markets, such as in the Netherlands where local partners have engaged to help Bing point users to reporting and support options for NCII.

For generative AI and multimedia, Bing 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 leverages its robust stack of mitigations and measurements to prevent the generation of nude or adult content.

Bing updated the report on 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 and harmful subjects.

As always and as described in the "How Bing monitors effectiveness of risk mitigation efforts" section above, Bing complements its targeted mitigations with continuous measurement, using both representative (realistic distribution of risky and non-risky queries) and adversarial datasets, as well as real-time social media monitoring, to detect emerging abuse patterns. As the service continues to evolve, Bing assesses how changes may affect systemic risks. For example, 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 support a mitigation strategy that remains proportionate, resilient, and aligned with user rights and expectations.

Generative experiences remain a meaningful driver of systemic risk focus, given their scale and the novelty of user interactions. During 2026, there were changes in the orchestration engine for Copilot Search and in the features and model supporting Image Creator and Video Creator, alongside iterative tuning to safety mitigations, to better manage user safety risks in generative contexts. These refinements operate within Bing's Safety Funnel, which provides a structured defense against harmful or misleading outcomes during model grounding, prompt handling, classification, and enforcement.

Year-Over-Year Progress in Risk Monitoring and Mitigation

Between 2025 and 2026, Bing's systemic risk posture remained broadly stable, with residual risk staying primarily in the Low range. Individual ratings moved little; where they shifted, the change reflected updated probability methodology — including updated user-reporting trends and the removal of social mentions as a scoring factor — rather than any fundamental change in Bing's service model or risk exposure. The year's substantive investments, described in "Changes to Bing's Risk Profile" above, included the expansion of the Defensive Defect Rate (DDR) program to track by DSA risk area, enhanced NCII detection and removal, expanded safeguards for generative experiences, and the formalization of product-development governance through pre-launch assessments, shiproom reviews, and Critical Impact Risk Assessments.

The combination of orchestration improvements and service-wide mitigations continues to shape Bing's risk profile by concentrating analysis on where safety systems interface with generative workflows (e.g., prompt enrichment, classifiers, and decisioning), and where defensive interventions can most effectively reduce residual risk.

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 2026 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.

The 2026 assessment reflects a complex global risk environment for VLOSEs, shaped by ongoing geopolitical instability, evolving information manipulation tactics, heightened regulatory expectations across regions, and the rapid expansion of GenAI capabilities for both beneficial and harmful purposes. While Bing's overall risk posture remains controlled, these shifts bring into sharper focus where residual risk is concentrated — most notably growing concern around NCII, synthetic intimate imagery, and other image-based abuse; continued attention to GenAI-enabled deepfakes and synthetic media; and a shift in civic-discourse and election-related risks following the 2024 global "year of elections" and continuing through the 2025–2026 period. Public concern in the EU around health misinformation appears lower than at the height of the COVID-19 pandemic, while broader mental- and physical-wellbeing risks remain relevant for sensitive queries, vulnerable users, and GenAI experiences, and violent-extremism and public-security risks remain present as specific actors and focal points evolve. These developments have not changed Bing's core function as a search engine, but they inform the 2026 assessment's more current, granular view of residual risk in Human Dignity, Illegal Content and Activities, and Private and Family Life.

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 (Coalition for Content Provenance and Authenticity) 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 support transparency in content creation and editing. Bing also implements the system of mitigations for Bing Gen AI experiences as discussed above. Based on available internal and external signals, Bing has not identified evidence that these risks disproportionately affect EU-based users or vulnerable groups through Bing's services, but Bing continues to monitor the issue as synthetic media risks evolve.

Operational orchestration has become a central risk-management theme for generative experiences, and Bing continued to apply targeted guardrails – model grounding to authoritative results, chat-turn limits, in-service transparency disclosures, hard/soft blocklists, and AI-based classifiers – with DDR tracking by DSA risk area remaining the cornerstone for assessing their effectiveness (see "Changes to Bing's Risk Profile" and the Generative AI Experiences Specific Mitigations section).

Changes in risk rating

Figure 14: Changes from 2025 to 2026 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

No changes have been noted in the Residual Risk Rating between 2025 and 2026.
Every risk area carries forward the rating it held in the prior assessment cycle, and the distribution across ratings is unchanged: one risk area rated Minimal, eight rated Low, and three rated Moderate.

For Residual Risk Rating, the number of three Moderates in 2026 is consistent with 2025. The three risk areas rated as Moderate Residual Risk — Human Dignity, Illegal Content and Activities, and Private and Family Life — are the same three areas rated Moderate in the prior year. These Moderate-rated areas reflect risks that are evolving, higher-impact, or subject to heightened public and regulatory attention, while remaining within an acceptable range given Bing's implemented mitigations and its unique risk profile as a search engine.

This year-over-year consistency should not be read as an absence of change beneath the ratings. Refinements were made to the probability methodology and inputs, including updated user-reporting trends and the removal of social mentions as a scoring factor. These changes resulted in a more current and conservative view of where residual risk is concentrated, while the underlying mitigation approach continues to rely on Bing's layered safety framework, targeted interventions, user reporting, DDR monitoring, and ongoing product governance. That the ratings held steady under a more conservative methodology reinforces, rather than undercuts, the assessed stability of Bing's risk profile.

What remains Moderate

Each of the three areas remained Moderate for reasons consistent with the prior year:

- **Human dignity** — the gravity of harms in this area, where mitigations are assessed as "Managed" but the Inherent Severity remains Critical.
- **Illegal content and activities** — the continued severity of illegal-content risks, including CSEAI/CSAM-related risks assigned to this area.
- **Private and family life** — continued concern around NCII and related image-based abuse risks, despite Bing's investment in NCII mitigations across products this year.

Bing assesses that current mitigations across these three areas are reasonable, proportionate, and effective, and will continue to prioritize them as it matures its capabilities across the safety funnel.

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 people 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 support algorithms that provide high quality, relevant content in response to the user's queries without unduly limiting their ability to access information or express themselves.

In 2025–2026, Bing continued to evolve its search and generative experiences while maintaining the same core safety stance, filters, and policies. These updates did not fundamentally change Bing's approach to balancing user safety with access to information. Bing continues to support media pluralism by preserving access to traditional search results and grounding generative experiences in authoritative sources where applicable.

Bing further supported this balance by bolstering its policy to evaluate both the user's intent and the potential risk associated with the topic for which they are seeking information. This framework enables Bing to apply more nuanced and targeted mitigations continuing to safeguard users' free

expression and rights to access information while managing risks of users unexpectedly encountering harmful, misleading, or offensive content.

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 apply narrowly tailored measures that do not unduly restrict fundamental rights. 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 provides 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 its 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 over=blocking rates (i.e., rates at which responses are restricted) to drive improvements. Bing also strengthened management of user safety risks in generative experiences: The program prioritized refining how generative responses are produced and governed, with specific orchestration and mitigation updates to improve safety outcomes.

Bing continues to preserve open access to information by maintaining traditional "blue links" alongside generative experiences. Generative summaries are grounded on search results and surfaced with appropriate citations, with ongoing work to support summaries that are high quality, safe, and not biased or leading to harm. These design choices help uphold freedom of expression and media plurality by contextualizing generative outputs and ensuring users can navigate to original sources.

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 over-blocking 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

Risk area	Freedom of expression and 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 searches 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 promote 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 area	Freedom of expression and information
	• Risk that GenAI protection systems flag and block the generation of legitimate content

Inherent risk

Freedom of Expression and Information are foundational considerations for search services, which often serve as a primary means for users to access and discover information online. By organizing and presenting vast amounts of information, search services inherently face the possibility that information may be restricted, unavailable, or surfaced in a manner that differs from user expectations.

As a result, 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 from 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 designed to help prevent 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 because GenAI features apply additional safety controls. However, this is 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 with freedom of access to information. This layered approach supports 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 respect 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 support 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 designed to support actions that 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. Generative answers remain grounded in search results and surfaced with appropriate citations to promote transparency, user agency, and media pluralism.

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) to continually assess the effectiveness and appropriateness of its GenAI safety measures to appropriately prioritize users' fundamental rights to information. Additionally, the Copilot Search vertical no longer includes image creation; that functionality is available separately on Bing, reducing the likelihood that visual content policies constrain expressive queries within Copilot Search while allowing tailored safeguards in the dedicated creative experiences.

Bing engages in regular self-directed tests and audits of its system to assess and support 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. Copilot Search does not modify the ranking of websites used for grounding, helping support the diversity of high authority sources reachable via search is preserved. Bing's core ranking approach remained stable during the assessment period, with incremental improvements focused on result quality and relevance. Generative experiences continue to rely on grounding, citations, and core search retrieval where applicable, supporting access to high-quality information and continuity with Bing's trust and safety principles.

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 assess whether Bing has sufficient protections in place to uphold free expression. Bing is bi-annually 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 people from unexpected exposure to content that could negatively affect their fundamental rights to human dignity. Bing provides users with a SafeSearch option to control their search experience to filter out explicit 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.

In 2025–2026, Bing's search and generative experiences continued to apply safety filters, classifiers, and policy restrictions designed to mitigate dignity-related harms, including hate, harassment, sexual content, graphic violence, and exploitative content. Orchestration and UX changes in Copilot Search did not introduce new capabilities that materially expanded human dignity risk, and some creative experiences such as Bing Image Creator and Bing Video Creator have been enabled based on more robust policy and safety interventions.

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
Risk scenarios	• 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 area	Human dignity
	• 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 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 content involving real or synthetic individuals without consent

Risk area	Human dignity
	• Risk that Bing Image or Video Creator is used to generate manipulate images that are violent, gory, vulgar, or include profanity • Risk that Generative AI summaries generate content that is graphic, violent, gory, vulgar, profane or that otherwise demeans, shocks, or offends • Risk that Generative AI summaries generate content that is sexual or pornographic or link to information on locating consensual and legal prostitution • 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, sexual exploitation, 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, as 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 1.21% from August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with human dignity harms and serves as one data

point to inform risk monitoring. This result is broadly consistent with the prior year's average DDR of 1.28%, with a slight decrease, indicating relatively stable exposure levels over time.

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 supports 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, so that 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 classifiers, in addition to threat intelligence, 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 prevent CSEAI from being 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 harm to human dignity and adjust the safeguards accordingly. Through this process, Microsoft enhances 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 explicit, violent, or gore results, NCII, 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 prohibits use of Bing's GenAI features to create or share explicit 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 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 and Activities on the service to be "Moderate" (Refer to *Changes in Risk Rating* above for further details). This Moderate rating is consistent with the 2025 assessment and reflects the continued severity of illegal-content risks, including CSEAI/CSAM-related risks assigned to this risk area.

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 and 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

Risk area	Illegal content and activities
	• Facilitation of unlawful immigration • Causing suffering of animals • Creation, dissemination, or promotion of CSEAI • Content or activities that otherwise disseminate 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 CSAM results are returned on News, maps, travel searches • Risk that Copilot Search or Bing Image or Video Creator are used to generate third-party IP without authorization in generated responses • Risk that Generative AI features generate content that violates local laws; 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 • Search Generative AI features could be used to create, locate, or trade in real or artificial CSEAI – content that is harmful even if the search-based tool did not generate it

Risk area	Illegal content and activities
	• Risk that Generative AI summaries inaccurately interpret source data to generate incorrect or misleading information about an individual that could be defamatory

Inherent risk

The **inherent probability** of systemic risks related to Illegal Content and Activities arising 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 August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with Illegal Content and Activities harms and serves as one data point to inform risk monitoring. This result is consistent with the prior year's average DDR of 0.14%, indicating stable exposure levels over time.

During the assessment period, Copilot Search maintained its existing safety filters, classifiers, and policy enforcement stance for illegal-content risks. Backend platform changes did not materially alter the applicable safety posture, and user reporting was improved by adding Copilot Search as a distinct category in the Report as a Concern experience.

Bing also advanced user reporting pathways relevant to this risk area. In April 2026, Copilot Search was added as its own category in the "Report a Concern" user experience, which is expected to improve granularity of user signal intake for safety teams as measurement and reporting mature.

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 supports 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.

Mitigation of illegal content and activity risks across the service is implemented through a layered approach. Ranking systems prioritize relevant, fresh, and authoritative results while incorporating safety, defensive, and spam signals. Harm-specific interventions, including PSAs, blocking determinations, and alternative ranking treatments, are managed by dedicated safety and defensive functions to help maintain consistent and policy-aligned handling of high-risk content.

Across generative search experiences, safety filters, and a conservative policy framework are applied to prevent the generation or amplification of illegal content and to limit pathways that could facilitate illegal activities. No material changes to these safeguards were introduced during the assessment period, reflecting a stable mitigation baseline for illegal-content risks.

Similarly, Bing also states in its policies that advertisers who are misusing Microsoft services in a way that could cause harm to individuals or the service, such as advertising in a 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 local legal experts.

Individuals 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 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 to protect privacy better. 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 identify a change in the scoring compared to the 2025 assessment.

Risk area	Inherent probability	Inherent severity	Inherent risk rating	Mitigation maturity	Residual risk rating
Protection of personal data	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 to 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 • Unexpected disclosure or exposure of personal information or data online • Other content or activities negatively impacting the protection of personal data

Risk area	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 GenAI features inaccurately interpreting source data to provide incorrect or misleading information about an individual • 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 "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. 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 "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 "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 August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with protection of personal data harms and serves as one data point to inform risk monitoring. This result is consistent with the prior year's average DDR of 0.02%, indicating stable exposure levels over time. 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 supports both platform-wide consistency and harm-specific responsiveness.

Bing data retention is managed within a set schedule. Bing limits the retention of data necessary to provide the service and helps protect the safety of the service. Bing's in-product control settings are provided to users through Bing's settings page; ***authenticated*** users can control and maintain their search history.

Copilot Search has data retention and deletion policies that limit retention of personal data collected through this feature to be 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 protect 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.

Access to most Bing Search history data is limited to Bing employees on a need-to-know basis and the specific user is de-identified. 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 submitted to Microsoft's Data Protection Officer and are subject to re-evaluation annually or sooner, should there be a significant change to processing risks. This process helps 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 individuals 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 individuals 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 election-related manipulated information, the potential for third-party content indexed in search or generated in AI features to contribute to political polarization, and targeted efforts by malicious actors to leverage technology to undermine civic institutions and democratic elections.

Search engines like Bing play a key role in helping people to 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, training, and support.
- Election authorities should be able to support 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 and 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 threat intelligence narratives related to election and information manipulation or hate 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 high risk 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".

Risk area	Inherent probability	Inherent severity	Inherent risk rating	Mitigation maturity	Residual risk rating
Civic discourse and electoral processes	Highly Likely	Critical	Critical	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 information manipulation • 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

Risk area	Civic discourse and electoral processes
	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 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 or misses filtering certain political merchandise containing false or misleading election-related claims • 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 that content promoting efforts to undermine election integrity, or mass activities of groups seeking to destabilize civic institutions (e.g., anti-democracy organizations) appears in Search results in Bing outputs • 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 area	Civic discourse and electoral processes
	• 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'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.

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. While the number of elections in Europe during the assessment period differed from the prior year, the overall probability assessment increased due to updated methodology inputs, including changes in user reporting data and removal of social mentions as a scoring factor.

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 "Highly Likely" and inherent severity as "Critical", the **inherent risk** associated with civic discourse and electoral processes on the Bing service is "Critical".

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" and 2025 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 August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with civic discourse and electoral processes harms and serves as one data point to inform risk monitoring. This result is consistent with the prior year's average DDR of 0.03%, indicating stable exposure levels over time.

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 supports both platform-wide consistency and harm-specific responsiveness.

Bing takes a multi-layered approach to protecting election integrity maintaining and continuously maturing a defensive playbook focused on elections, 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 supports all advertising on Bing. Microsoft Advertising policies prohibit ads for election-related content, including political candidates, parties, ballot measures, and political fundraising globally. Ads supporting fundraising for political candidates, parties, political action committees, or ballot measures are also prohibited. Microsoft Advertising policies also prohibit issue-based advertisements, as defined in its published policy. Microsoft Advertising is a signatory to the EU Code of Conduct on Disinformation (COCD) under the DSA , and further details on its commitments are set out in Microsoft's most recent COCD report.

Over the past year, Microsoft Advertising maintained this prohibition and introduced additional compliance measures in alignment with EU Regulation (EU) 2024/900 on the Transparency and Targeting of Political Advertising (TTPA), which entered full application on 10th October 2025. The definition of political advertising applied to ads serving the European Union was updated to align with the scope set out in the regulation.

To enforce this prohibition at an earlier stage in the campaign lifecycle, a mandatory advertiser Political Advertising Declaration was introduced in September 2025. Advertisers must declare, before campaign creation or import, whether a campaign contains political advertising. Because Microsoft Advertising does not permit political advertising, any campaign declared as political is

blocked from serving. This establishes a proactive, declaration-based control at the point of campaign entry, rather than relying solely on downstream detection. Non-compliance, including misrepresentation of a campaign's political nature, can result in account suspension or termination. Campaigns active prior to these changes remain subject to ongoing review to confirm continued compliance.

Product safety

Beyond standard measurement to evaluate the quality of search results against our policies, and ahead of specific high-risk 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 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).

In addition, through paid partnerships with fact checkers, researchers and proprietary pipelines, Bing actively tracks manipulated or harmful narratives related to national elections and civic integrity matters in EU member states and apply threat-intelligence informed defensive search interventions to reduce the visibility of low-quality content on related queries. Additional to our ranking interventions, Bing may temporarily deploy ***special informational and answer segments*** *for select high risk* 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.

Generative search features such as Copilot Search and Bing Image and Video Creator utilize defensive interventions such as extra safeguards to grounding and citations on high-risk queries, as well as local election mitigations, such as blocklists, to restrict generation of images or certain types of content concerning active candidates in upcoming high-risk elections or to limit responses to certain election-related queries out of an abundance of caution. Bing continues to invest in these mitigations, including through query classifiers, focused testing ahead of high-risk elections and threat intelligence intake 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***, supports Microsoft 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 to support the cybersecurity maturity and resilience of democratic stakeholders through nonprofit partnerships. Microsoft has worked with IFES, International IDEA, UNDP, NDI, IRI, Freedom House, Witness, and others to provide programs and trainings in service of journalists, elections officials, political parties, observers, human rights defenders and more.

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.

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 COCD*** and as a signatory to the EU COCD, 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 people 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 COCD framework.

As part of its commitments under the EU Code of Conduct on Disinformation (COCD), 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 COCD 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 COCD.

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 individuals 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 have been a major focus for regulators with concerns about how GenAI exacerbates these risks top of mind. Importantly, while Bing assesses and manages risk associated with fraud, search engines do not have the same risk as platforms that push and enable sharing of content.

While GenAI introduces new vectors for fraud-related misuse, Bing's existing layered mitigations, including advertising controls, ranking systems, and classifier-based safeguards, continue to effectively limit exposure to these risks, supporting a "Low" Residual Risk Rating. 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
Civic discourse and electoral processes	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 to occur on the service.
Risk scenarios	• Scamming, defrauding, spamming, use of malware, virus, or spyware, fraudulent business or advertising, or identity theft

Risk area	Consumer protection and fraud
	• 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 advertisements 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 of 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 Generative AI summaries generate ungrounded results or mis-characterize cited web links that lead to misinformation about the legitimacy of businesses, products, or services • Risk that GenAI summaries are based on low authoritative sources and produce content that may be harmful or misleading

Risk area	Consumer protection and fraud
	• 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 from 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 is 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 August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with consumer protection and fraud harms and serves as one data point to inform risk monitoring. This result is consistent with the prior year's average DDR of 0.02%, indicating stable exposure levels over time.

Bing continues to apply managed, policy- and classifier-driven safeguards to reduce exposure to scams, deceptive or fraudulent content, and misleading commercial practices across search results and generative surfaces. 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 supports 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 provide users with high-quality content to avoid the possibility 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 also, including investments in threat intelligence and social listening in the period and deploys mitigations to support the return of quality results to users. Continuation of Bing's broader content governance and enforcement mechanisms (framework-level policies, monitoring, and incident response) that address deceptive practices and spam across surfaces, supporting a "Managed" mitigation maturity and "Low" residual risk assessment for this risk area.

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***.

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. Policy and classifier gating to detect and constrain responses associated with scams, fraud, and other deceptive or misleading commercial practices, including limits/refusals where content would violate safety policies

The ***Microsoft Advertising Fraud Detection System*** is multi-tiered, combining machine-learning (ML) models, heuristic assessments and rules, and manual review to prevent and detect fraudulent ads from entering the system, with models and rules that apply both at the advertiser level and to individual ads.

Ad-level models are applied based on the Microsoft Advertising Policies relevant to the market where the ad serves and include, among others, CSEAI, offensive/hate speech, violence, suffering, gore, exploitation, explicit content, gambling, tobacco, weapons, and prohibited and restricted healthcare and products.

Advertiser-level controls and models include Advertiser Identity Verification (AIV), mandated for advertising customer accounts, which requires advertisers to prove their legal identity before running ads and is reinforced with domain-verification checks (including the Domain Verification Tool), which confirm domain ownership and registration. These control are a deterrent to phishing, personal-information fraud, and disallowed-product advertising, payment fraud detection, a purchase risk model (triggered at billing to determine whether to attempt the charge or block the advertiser), a behavioral risk model (evaluating ad text, keywords, budgets, account age and structure, payment-instrument attributes, associated individuals, and access devices whenever an

advertiser changes campaigns), and LLM-based spoofing detection targeting advertisers impersonating legitimate sites. Building on these, the system now also includes LLM-based celebrity-scam and brand-impersonation detection (multilingual models that identify false celebrity endorsements, brand and company impersonation and phishing, and counterfeit patterns, reducing scam impressions), account-compromise (account-takeover) detection, which were introduced in response to a rise in account compromises to overall compromise impact, and an autonomous AI review agent that support domain review processes and enables automated enforcement actions in defined scenarios.

Risk detection is organized through a structured fraud taxonomy of "constellations" (groupings of related fraud patterns used to index, escalate, and report complex scenarios) and "patterns" (granular tactics such as cloaking, phishing, and baiting), supported by an internal "MO tracker" that captures suspicious accounts not matching known patterns. The classifiers and rule-based systems evaluate behavioral signals (spend rate, budget/bid patterns), identity signals (account, customer, payment-instrument, device), text and visual content (ad text, images, landing-page screenshots), and link signals (redirects, cloaking, lookalike domains), with heightened scrutiny of high-severity-harm traits such as guaranteed-return claims, violations of regulatory requirements such as disclaimers or risk warnings, urgent financial calls to action, newly created or rapidly scaled accounts, repeated violations, payment anomalies, and cloaking.

Based on model output, Microsoft Advertising may take enforcement action, including account suspension. The system also incorporates internal and external signals. Examples of internal signals include a Bing signal indicating whether a domain carries spam and a Microsoft Defender signal indicating user harm from visiting a page. Examples of external signals include aggregated malware-risk signals from VirusTotal, domain-risk tools, domain age/expiry, and cross-engine click data from Whois.

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.

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.

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 impact 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 for 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 contains 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 area	Discrimination and hate
	• Risk that search engines return content that provides hateful rhetoric or advocates for 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 perpetuates bias, reinforces stereotypes, or discriminates 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 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; risk can be increased if prompts do 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 area	Discrimination and hate
	• 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 from sufficient mitigations is assessed as "Not Likely".

Bing considered internal metrics and external data as key factors in assessing a "Not likely" level of probability relative to other potential systemic risks on Bing, reflecting current observed risk patterns rather than a year-over-year reduction.

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 .13% from August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with discrimination and hate and serves as one data point to inform risk monitoring. This result is broadly consistent with the prior year's average DDR of 0.14%, with a slight decrease, indicating stable exposure levels over time. 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 supports 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 to providing users access to high-quality and authoritative content online, Bing's ranking systems evaluate objective quality and trustworthiness signals. Content exhibiting indicators of low-quality or potentially harmful content, including harassment or derogatory language directed at individuals or protected groups, may be assessed as lower quality and receive less prominent placement in search results.

For generative experiences, Bing continues to apply additional mitigations to reduce the risk of biased, discriminatory, or hateful outputs. These include grounding generative responses in high-authority, trustworthy web content; applying classifiers and content filters; and continuing work on grounding metrics and safety gen practices intended to improve the quality, safety, and citation practices of AI-generated summaries. These controls are relevant to discrimination and hate because they reduce the likelihood that generative summaries will rely on low-authority sources, reflect harmful stereotypes, or present discriminatory content without appropriate mitigation.

The Terms and Conditions for Users, found within the Microsoft Services Agreement, prohibit using Microsoft services to engage in activity that is harmful to others, including communicating hate speech. The Bing Image Creator Terms of Use further prohibits use of Bing's GenAI features to create or share hateful content or content intended to exploit the vulnerabilities of individuals or groups.

Bing has a ***Threat Intelligence*** team that is accountable for implementing algorithmic interventions and metrics monitoring of viral and likely manipulated narratives and conversion dedicated to identifying and remedying, via targeted algorithmic interventions. These interventions are designed to reduce exposure to 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. In the period, Bing has further invested into partnering with threat research organizations in the broader domain.

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, so that 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 also expanded external coverage through partnerships with Associated Press (AP), the Institute for Strategic Dialogue (ISD), and Lead Stories. These partnerships enhance detection and response to emerging harms and provide additional support to the fact-checking ecosystem that underpins authoritative sourcing in search results

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 the 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	Likely	Critical	High	Optimized	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 • Addictive 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

Risk area	Mental and Physical Wellbeing
Theoretical Risk Manifestations absent Mitigations	• Risk that search results or autosuggest 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, promotes, normalizes, praises, or facilitates suicide or self-harm. • Risk that or 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 oversimplified Generative AI summaries generate 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 is in the top search results. • Risk that Shopping listings show products with logos or graphics that promote or 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 harm. • 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 area	Mental and Physical Wellbeing
	• 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 from 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 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 "Likely" and Inherent Severity as "Critical", the **Inherent Risk** associated with negative impacts to Mental and Physical Wellbeing on the Bing service is "High".

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 Residual Risk Rating for mental and physical wellbeing decreased from Moderate to Low in 2026, due to the implementation of a comprehensive and continuously evolving 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 the risk-specific mitigations summarized below. These mitigations operate across multiple layers of the service, including product design, content moderation, targeted interventions, incident response, and external engagement.

The Mitigation Maturity for Mental and Physical Wellbeing is rated as "Optimized". This rating reflects Bing's mature and continually improving approach to identifying, mitigating, monitoring, and responding to risks associated with gender-based violence, harassment, suicide and self-harm content, eating disorders, and other threats to users' mental and physical wellbeing. Bing has supplemented its longstanding safety controls with additional targeted interventions, operational readiness measures, and partnerships that strengthen both prevention and response capabilities.

During the assessment period, Bing expanded its Threat Intelligence capabilities through the introduction of a dedicated Events Playbook and enhanced collaboration with internal and external partners to identify and manage emerging risks. These capabilities improve Bing's ability to monitor, assess, and respond to rapidly developing safety risks, including those affecting people's wellbeing.

Bing also enhanced its user protection measures through expanded PSAs and defensive search interventions. These include specialized user-support experiences and authoritative-resource pathways for queries related to NCII, eating disorders, and suicide or self-harm in specific markets. These interventions are designed to reduce exposure to harmful content, promote authoritative information, and direct users to appropriate support resources when engaging with high-risk queries.

In addition, Bing continues to leverage authority-based ranking systems, targeted algorithmic interventions, content policies, AI safety controls, user reporting mechanisms, social listening, and ongoing monitoring of safety metrics to support user wellbeing across both traditional search and GenAI-enabled experiences. Collectively, these measures demonstrate a mature mitigation framework that is regularly evaluated and enhanced based on emerging risks, operational learnings, and external stakeholder engagement.

While not a proxy for Residual Risk, Bing has measured the DDR for Mental and Physical Wellbeing on Bing Search as an average of 0.06% from August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with Mental and Physical Wellbeing harms and serves as one data point to inform risk monitoring. This result is broadly consistent with the prior year's average DDR of 0.07%, with a slight decrease, indicating stable exposure levels over time. Bing continues to monitor safety indicators, user reports, product signals, and emerging risk trends to evaluate the effectiveness of existing mitigations and identify opportunities for further enhancement.

As a result of these enhancements and the breadth of controls deployed across Bing's ecosystem, the mitigation framework for Mental and Physical Wellbeing is assessed as sufficiently mature to reduce the residual risk associated with this risk area to **Low** despite the continued assessment that the underlying inherent risk remains **High** due to the seriousness of potential harms and the widespread prevalence of related content and behaviors online.

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 supports 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 reduce the risk that users are not offended, harmed, or misled by problematic material in search results.

Bing provides additional context to help people avoid harm from content surfaced in search results and may ***point to authoritative information***, when search results lack high authority content, or redirect users when they search for specific queries. For example, when people 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 of self-harm, self-harm images, or videos. Queries related to self-harm or suicide are considered high-risk queries, and harmful content associated with these queries are considered low quality and generally demoted. While demoting user-generated content is tricky for image and videos due to quality control signals from social media, PSAs redirect them to help lines. Similarly, PSAs are also shown for user search queries potentially related to searching for Child Sexual Abuse Material, directing people to hotlines and reporting resources. In Copilot Search, self-harm and suicide-related queries are treated as high-risk. In response, Copilot Search responds to those queries through "thoughtful disengagement" and directs people 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 people 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 (the nonconsensual disclosure of an individual's personal information), 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 remain a heightened area of focus, including concerns about GenAI-enabled creation or amplification of NCII and user access to sites, tools, or applications that enable the creation or modification of intimate imagery, including so-called "nudify" applications.

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 risk queries and as appropriate, remove potentially harmful content related to NCII and reduce the visibility of nudify

apps as low-quality content. Furthermore, Bing uses safeguard search features such as moderating search suggestions, and limiting or safeguarding generative summarizations on identified NCII related queries to reduce the risk of leading users to NCII content, tooling or instructions. Additionally, Bing has started to implement public safety messages in the search results page such as in the Netherlands where Bing received engagement from local partners, pointing users to reporting and support options for NCII.

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 maintain effectiveness without compromising user access to information, as referenced above.

Mitigations and controls remain materially consistent with prior year. The same Responsible AI response classifier gates prompts and responses under the current serving architecture, applying unchanged policy restrictions to block or constrain sensitive or harmful content that could implicate private and family life interests. Backend serving and model upgrades delivered measured improvements in defensive defect rate (DDR). While framed as technology improvements rather than feature changes, a lower DDR is consistent with fewer policy-violating responses, including those that could intrude upon private life or disclose personal data.

Private and family life's Residual Risk Rating increased from the prior year assessment due to an 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).

Private and Family Life remained Moderate in 2026, consistent with 2025, reflecting continued concern around NCII and related image-based abuse risks, as well as the assessment that the mitigation maturity for this area does not yet support a lower Residual Risk Rating.

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

Risk area	Private and family life
	• 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, where required, 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 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 from 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 August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with Private and Family Life harms and serves as one data point to inform risk monitoring. This result is consistent with the prior year's average DDR of 0.02%, indicating stable exposure levels over time.

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 supports 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 services to violate or infringe upon the rights of others or engage in activities that violate the privacy of others. Although Bing Search generally does not allow users to publish content directly to the service, these terms help establish expectations for the lawful and responsible use of Microsoft services. 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 (i.e., 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 including 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 help users distinguish AI-generated content from authentic media and reduce the risk of deception or misrepresentation. 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 assess whether 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 helps embed safety, privacy, and integrity 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 explicit 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 prevent the material from appearing 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 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 NCII, 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 facilitate 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 people 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,

Risk area	Public Health
	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 bad actors intentionally manipulate search results to increase the visibility of health misinformation, including outdated guidance on regulated goods. Risk that Generative AI 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 Generative Search and produce harmful or misleading content.

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 from 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 August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with Public Health harms and serves as one data point to inform risk monitoring. This result is consistent with the prior year's average DDR of 0.04%, indicating stable exposure levels over time.

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 supports both platform-wide consistency and harm-specific responsiveness. For authoritative grounding for health-related queries, Bing emphasizes grounding answers in trusted, high-authority sources and elevating vetted domains to improve reliability. Health responses are only generated using ***grounding data from trusted health domains***, which have been vetted by the Bing Health team.

For risky-query identification and presentation controls, Bing is investing in better identification of queries that may carry elevated Public Health risk (e.g., those likely to trigger harmful or misleading health narratives) and applying stricter authority thresholds, source citation practices, and presentation controls for those cases. These efforts focus on maintaining reliability as generative experiences become a more common way people consume information in search.

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 people to trusted sources through several tools. For events such as select elevated risk EU elections that may impact an exceptionally high number of users, in addition to algorithmic search interventions on threats, Bing may highlight answers that point to expert-vetted information. During events like emerging wars, violent events or 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 extremism 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 combating 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 COCD 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 area	Public security
	• 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 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 area	Public security
	• 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 Generative AI summaries generate ungrounded results or mischaracterize cited web links that lead to misinformation about public safety incidents. • Risk that Generative AI 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 from 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 .06% from August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with Public Security harms and serves as one data point to inform risk monitoring. This result is broadly consistent with the prior year's average DDR of 0.07%, with a slight decrease, indicating stable exposure levels over time.

Expectations among EU authorities have shifted toward demonstrable, measurable effectiveness of mitigations, with substantive scrutiny of methodology and evidence supporting risk-reduction claims. Bing is accordingly emphasizing clearer measurement frameworks and traceable evidentiary support for how interventions reduce Public Security risk across relevant surfaces.

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 supports 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 41 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 individuals from harmful web content, Bing offers a SafeSearch feature to manage exposure to explicit content. SafeSearch is available without authentication. Additionally, Microsoft Family Safety features enable parents to manage and track use of applications, including Bing.

Bing does not require people to authenticate to access search. This is standard practice across search engines and is vital given the role that search engines play in enabling individuals to freely access information. Allowing people to search without authentication is also an important privacy consideration for both adults and minors as authentication requires the submission of personal information. In the EU, Bing defaults to SafeSearch Moderate mode for unauthenticated users; SafeSearch Strict mode filters explicit content; however, Moderate mode provides blue links in response to requests for adult content. Users who are not identified through authentication as minors are treated as adults.

Increasing Age Assurance in the Search Context

Bing recognizes the increasing focus on providing age-appropriate online experiences for children and teenagers. Approaches to age assurance continue to evolve globally, reflecting differing legal frameworks, user expectations, technological capabilities, privacy considerations, and regulatory requirements. As a search service, Bing differs from services centered on sustained engagement. People primarily interact with Bing to seek information or complete discrete tasks.

Search services historically only knew users' ages if those users chose to provide that information, for example, by self-declaring their age while setting up a Microsoft Services Account and choosing to sign into Bing. The majority of Bing users are unauthenticated, and Bing had limited visibility into their ages, which creates challenges in delivering age-appropriate experiences while maintaining access to information and respecting user privacy.

Bing continues to evaluate and test approaches that may improve protections for younger individuals and monitors developments across industry, regulatory, academic, and civil society efforts as age assurance technologies and practices mature.

Bing remains a search service and does not require authentication for general search access. Bing is actively evaluating how age-related signals can be used in a privacy-preserving and proportionate manner to strengthen protections for minors. Bing's approach reflects the need to balance child safety with privacy, data minimization, access to information, and jurisdiction-specific requirements.

As part of this ongoing work, Bing is taking steps to strengthen SafeSearch protections for users determined to be under 18, including defaulting such users to Strict SafeSearch. Bing is also monitoring broader Microsoft work related to common identity infrastructure, including Microsoft Account-based age signals, where such signals may support more consistent and privacy-conscious protections across consumer experiences. These efforts are product-specific and geography-specific, and Bing will continue to assess how age assurance approaches can be applied appropriately in the search context.

In assessing risks to minors, Bing also considered the European Commission's child safety guidance and the 5C typology as additional analytical lenses for identifying how harms to children may arise through content, contact, conduct, consumer harms, and cross-cutting risks. These frameworks do not replace Bing's DSA-aligned risk areas, but they help inform how minor-related risks may manifest across Core Search, generative AI experiences, and ancillary search features.

Microsoft believes there is value in minors engaging with GenAI services for legitimate purposes, including learning, creativity, and information. 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 restrictions.

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".

In assessing risks related to the Rights and Protections of Minors, Bing also considered the 5C typology as a complementary framework for understanding how children may experience harm across digital services. This typology supported Bing's evaluation of minor-related risks by helping to organize harms associated with exposure to harmful content, harmful interactions or contact, harmful conduct, consumer-related harms, and broader cross-cutting risks that may affect minors' safety and development. Applying this lens reinforced Bing's understanding that risks to minors may arise differently across search, generative AI experiences, and ancillary features, and that mitigations must therefore be tailored to the specific ways such harms may manifest.

5C typology for protection of minors

The 5C Typology refers to the minors-risk framework covering five dimensions:

1. *Content harms*
2. *Contact harms*
3. *Conduct harms*
4. *Consumer harms*
5. *Cross-cutting harms*

The use of the 5C typology also supported a more structured consideration of how Bing's existing mitigations, including SafeSearch, ranking and authority signals, product safeguards, targeted interventions, content moderation, and incident response, contribute to reducing risks to minors across the service.

In alignment with the EU 5C Typology for the protection of minors, Bing considers risks across five dimensions: content, contact, conduct, consumer, and cross-cutting harms. These dimensions inform how risks to minors are identified and mitigated across Bing's services.

Risk area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
Rights and protections of minors	Not Likely	Critical	Moderate	Managed	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 to 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

Risk area	Rights and Protection 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. [5C - Contact and Conduct] • 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. [5C – Contact and Consumer] • 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. [5C – Contact and Consumer] • Risk that Bing's data processes harm minors or collect its data without parental consent. [5C – Consumer and Cross-cutting] • 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, explicit content, information promoting or soliciting illegal and regulated goods, self-harm or content inciting violence against others. [5C – Content and Cross-cutting] • Risk that search engines return content that shows risky behaviors to minors, such as dangerous challenges in the top search results. [5C – Content and Conduct] • Risk that auto-suggested searches expose minors to content that perpetuates discrimination and cyberbullying. [5C – Content and Conduct] • Risk that search engines return content to minors that normalize eating disorders in the top search results. [5C – Content] • Risk that Bing's safety systems for search unduly limit minors' access to information. [5C – Cross-cutting] • Risk that Advertisements on Bing expose minors to harmful activities targeted for child sexual exploitation and abuse. [5C – Contact and Consumer] Risk that Shopping, Maps, News, and Ads may harm minor users by collecting minor data without parental consent and target children with minors with personalized content. [5C – Consumer and Cross-cutting]

Risk area	Rights and Protection of Minors
	• Risk that Advertisements on Bing target minors. [5C – Consumer] • Risk that rental listings could be used to facilitate child sexual exploitation. • Risk that Shopping, News, and Ads results expose minors to harmful third-party content such as hate speech, extremist content, violent or graphic material, explicit 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. [5C – Content and Consumer] • Risk that those age-inappropriate activities are recommended for children and/or families in Travel guides, or on the Maps. [5C – Content and Consumer] • Risk that Shopping or News could expose minors to products or content that perpetuate discrimination. [5C – Content and Consumer] • Risk that Bing's safety systems for Ancillary Features unduly limit minors' access to information. [5C – Cross-cutting] • Risk that images of children which do not conceal their privacy are shown on the "Travel the World Virtually" or in the Maps features. [5C – Contact and Consumer] • Risk that Bing Image or Video Creator generates harmful content inappropriate of/for minors [5C – Content and Cross-cutting] • Risk that Bing Image or Video Creator minor user prompts could be stored or collected without parental consent. [5C – Consumer] • Risk that Bing Image or Video Creator is used to create imagery that glorifies and promotes self-harm for minors. [5C - Content and Conduct] • Risk that Bing Image or Video Creator promotes risky behaviors among minors, such as dangerous challenges or trends, through compelling images • Risk that users could use generative AI search tools or features to test grooming tactics.

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 from 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 "Managed", which brings the **Residual Risk** Rating down to "Low".

The Mitigation Maturity rating of "Managed" 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 .49% from August 2025 to May 2026. This metric estimates the proportion of search traffic that may be exposed to content associated with Rights and Protections of Minors harms and serves as one data point to inform risk monitoring. This result is lower than the prior year's average DDR of 0.54%, indicating a modest reduction in measured exposure over time.

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 supports 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 individuals, 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 where the user is known to be 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, including Bing, ***does not deliver personalized advertising to children*** whose birthdate in their Microsoft account identifies them as under 18 years of age, as stated further in the Microsoft's Privacy Statement on children and advertising. Across Bing features, known authenticated users under 18 do not receive targeted advertising, and this approach is applied to new Bing features as they are introduced. Age-targeting segments for users under 18 are not offered within Microsoft Advertising. Under the Microsoft Privacy and Personalization policy, advertisers must not (i) create a remarketing list, retarget, or otherwise profile any individual under the age of 18; (ii) target individuals under the minimum age required for the product advertised; or (iii) implement remarketing on sites or applications directed to minors under 18 or other applicable age limitations in a given market. Microsoft also maintains clear policies prohibiting behaviorally targeted ads, or ads containing adult material such as alcohol or gambling, from being shown to users identified as underage, and uses both manual and systematic methods to review ads for compliance, reserving the right to remove non-compliant content. These protections are enforced both contractually and through the absence of under-18 targeting options in the advertising manager.

For ***pre-launch risk assessments***, regarding minors, there are specific standards that must be adhered to. For example, privacy reviews aim that minor users are defaulted to the highest 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 applies Microsoft policies and data use requirements to minor users, and considers 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 explicit 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 individuals 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 facilitate 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 services to engage in any activity that exploits, harms, or threatens to harm children.

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 people 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 support any new features that may implicate user content are proactively reviewed, and additional risk mitigations are 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 the 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 maintain 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 shared by internal stakeholders during risk assessment and mitigation activities is handled in accordance with Microsoft privacy and security standards to maintain 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 noted above, 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 2025 Systemic Risk Assessment report, Bing outlined five primary areas for enhancement of mitigations during the initial assessment period. Bing has made enhancements in each identified area as described below.

Enhancing governance and pre-launch review.

Advancement: Enhanced risk management and governance

Bing continued to enhance its governance framework for systemic risk management, including strengthening product governance processes and further integrating risk assessment activities into product development and oversight practices. Bing also continued to assess and manage risks in alignment with the Digital Services Act, including incorporating the Article 28(5)(c) Topology into its risk management activities.

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

Advancement: Advanced trustworthy search principles

Bing further matured its multi-dimensional policy framework supporting implementation of the Bing Trustworthy Search Principles across experiences. These enhancements strengthened the consistency with which safety, quality, and ranking interventions are applied across Bing, helping balance access to lawful information, user safety, freedom of expression, and other fundamental rights while adapting to evolving technologies and risks.

Continuing investments in targeted interventions to better manage systemic risks.

Advancement: Detected and responded to emerging risks

Bing continued to invest in its ability to detect and manage dynamic and complex risks. This included ongoing enhancements to threat detection and risk management capabilities, adaptation of mitigations to address emerging harms such as risks associated with so-called nudification applications, and continued expansion of Public Service Announcements (PSAs) to support vulnerable individuals and provide access to relevant resources and assistance.

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

Advancement: Expanded Collaboration with Partners and Continuous Improvement
Bing strengthened its incident response and operational monitoring capabilities through enhanced tracking of operational metrics, including action rates and turnaround times for incident response and limited content moderation activities. Bing also introduced an Events Playbook to support preparedness for and response to emerging events and rapidly evolving risk scenarios, complementing existing Threat Intelligence capabilities.

Improving children's safety.

Advancement: Improved children's safety through enhanced SafeSearch protections

Consistent with its commitment to protecting younger users, Bing enhanced child safety protections during the assessment period. Notably, Bing updated SafeSearch globally to prevent explicit image and video content from appearing in Moderate mode and shifted known minor users in the EU to Strict mode by default. These changes were designed to reduce the likelihood that younger users encounter harmful or age-inappropriate content while preserving access to lawful information.

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.

Launching a continuous monitoring dashboard.

Bing will bring pre-launch and post-launch metrics together into a single continuous monitoring dashboard. By consolidating signals that today sit across separate review and measurement processes, Bing intends to strengthen its ability to observe mitigation performance over time, support regulatory tracking and reporting obligations, and identify opportunities for improvement earlier and with greater confidence.

Continuing investments in threat intelligence and targeted interventions.

Bing will continue to invest in refining and enhancing its limited, targeted interventions designed to manage systemic risks effectively. Over the coming period, Bing will focus in particular on advancing its threat intelligence and intervention capabilities related to NCII and fraud, so that its mitigations continue to respond to changes in the broader ecosystem and to the evolving tactics of malicious actors.

Deepening engagement with the broader ecosystem.

Bing will continue to engage internally as well as externally with regulators, civil society, researchers, industry peers, and other stakeholders to deepen shared understanding of the unique risks and challenges that apply to search, and to inform the design of mitigations that are proportionate and effective in that context.

Updating Bing's Risk Assessment Workbook.

Bing will update the Risk Assessment Workbook that supports its SRA, including by streamlining how mitigations and the mitigation maturity assessment are captured. The objective is to enable greater clarity on Bing's layered system of mitigations — how individual measures reinforce one another — and to make changes over time more visible and easier to evaluate across assessment periods.

Better aligning regulatory cadences with core business processes.

Bing is working to align its regulatory cadences more closely with core business planning and review processes. The aim is to evolve from a once-annual exercise — supplemented by intermittent CIRAs as required — to a regularized, routine risk assessment and audit process that is recognized and rewarded within the business, embedding risk assessment into how teams plan, build, and ship.

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 to continue refining 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 facilitate 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 facilitate 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 to 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.

#	Risk area	Risk definition
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 to 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 to 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 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 users' 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 of 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 support 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 SRA.

- 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 and 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 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 15: 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 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 harm.

Figure 16: 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 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 17: 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 18: 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 aid 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.

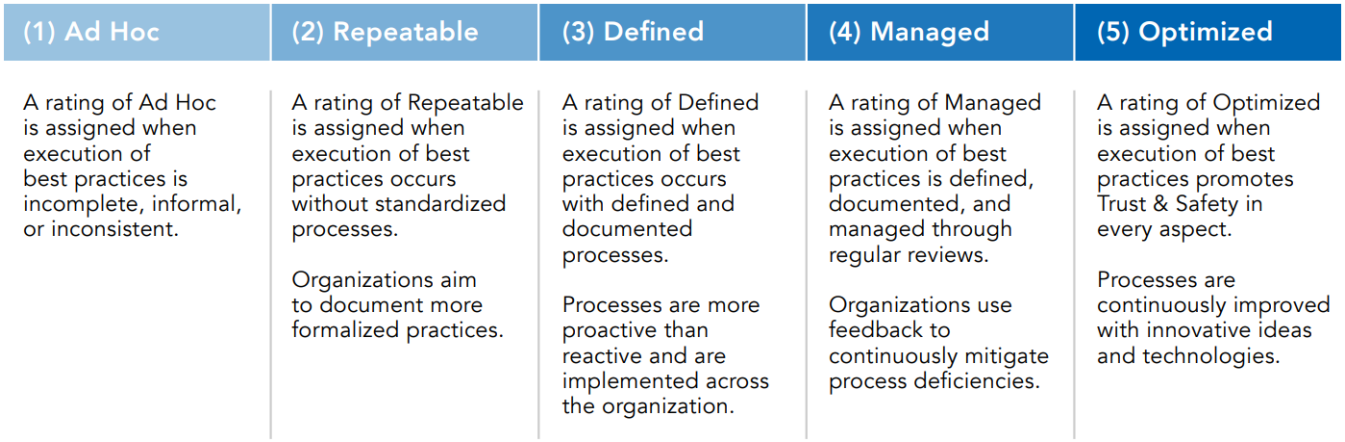
Bing stakeholder's complete product specific information collection sheets to identify implemented best practices within the listed Bing Risk and Compliance Framework best practice areas. Bing does not only develop and implement mitigations aligned with this framework; rather, the framework serves as a useful assessment to identify areas in which Bing can continue to mature implement mitigations.

Figure 19: Bing Risk and Compliance Framework Best Practices Screenshot

Governance & Oversight			Product/Service Design and Implementation				External Actions
Governance	**Compliance Management**	**Risk Management**	**Responsible Product/ Service Design**	**Policies & Other Mechanisms to Counter Harms**	**Content & Activity Moderation**	**Data Governance**	**External Engagement**
Governance & Oversight Convene governance forum, Senior Management oversight function	**Compliance Policies, Frameworks & Enablers** Define internal governance policies and operating models	**Risk Assessment & Identification** Conduct programmatic and systemic risk assessment(s), identify and prioritize risk	**Product / Service Development Lifecycle** Conduct safety and compliance reviews prior to product or service launch and throughout development lifecycle	**Policies & Policy Development** Develop policies for unacceptable service/product use and response to violative content/actions	**Policy Enforcement Mechanisms & Infrastructure** Establish activities to monitor and address violative content/activity with transparency	**Data Management Lifecyle** Define ownership, handling, retention, removal, and destruction protocols	**Public Reporting and Transparency** Publish transparency reports including ads repository, Systemic Risk Report, SOR Database
Roles and Responsibilities Determine single points of contact, legal rep, and management body / board oversight	**Communication & Training** Communicate and educate internal staff on roles, responsibilities, and requirements	**Risk Decisioning & Mitigation** Determine risk tolerance, decision and mitigation pathways, and escalation	**Product Features and Functions** Design products (algos, recommenders) with functional / transparency requirements in mind and limit platform misuse	**Terms & Conditions** Ensure application of fair, reasonable, and non-discriminatory general terms and conditions of access for users	**User Reporting** Enable and effectively address user reporting of violative content /activity	**Privacy, Access, and Acceptable Use** Adopt baseline controls on user access and data portability, enable third party access to data	**Public & Civil Society Engagement** Publish annual transparency reports, consult with civil society on systemic risk and impact, share data with researchers
Planning & Resourcing Determining adequate resourcing across the firm to meet obligations	**Compliance Assurance** Conduct controls testing, refer remediation to relevant teams	**Risk Monitoring & Reporting** Establish continuous monitoring and report on changes in risk status	**Responsible AI** Policies, procedures, processes and tools to help ensure responsible use, monitoring and maintenance of data and AI	**User Education and Empowerment** Inform and empower users on safe platform use, provide user choices and system optionality	**Appeals & Complaints** Establish complaint and redress mechanisms and out of court dispute settlement	**User Choice & Transparency** Adopt baseline controls that helps users build autonomy over how their data is used and processed	**Regulatory Response & Engagement** Satisfy external reporting requirements, respond to data access requests
Regulatory Convergence & Change Management Integrate upcoming requirements into existing compliance infrastructure	**Issues / Incident Management & Monitoring** Validate that program enhancements proceed timely; escalate to governance forums	**Crisis Response** Develop and implement protocols for addressing crisis situations					**External Audit** Facilitate external audit and refer remediation to 1 LOD product teams
Fees and Admin Requirements Pay applicable supervisory fees, purchase insurance and other admin requirements							**External Mitigation** Coordinate and share data with national authorities as appropriate

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 20: DTSP Maturity Rating Scale screenshot



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 21: 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

Description	Risk coverage	Framework coverage	Final score	Rating
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 22: 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

Description	Score (>)	Score (<=)	Score (<=)
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 one minus mitigation to achieve residual risk.

Figure 23: Systemic Risk Scoring Equation Screenshot



5C Typology

As part of the detailed methodology supporting this year's assessment, Bing incorporated the 5C typology as a supplementary framework to inform the identification and organization of minors-related risks. The typology was used to map potential harms to children across Bing's service surfaces and to support the analysis of how such harms may present within the report's broader DSA-aligned risk areas. In practice, this meant considering whether a given minors-related risk was primarily associated with harmful content, harmful contact, harmful conduct, consumer-related harms, or broader cross-cutting harms, while also recognizing that some risks may span more than one category. This supplemental mapping supported consistency in risk discovery, improved traceability between identified harms and relevant mitigations, and provided an additional external reference point for the assessment of risks affecting minors.

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. While mitigations may address more than one component of the Bing Safety Funnel, for clarity and ease of review, each mitigation is presented in ***bold italics***, followed by a brief description of how the mitigation supports the relevant Bing Safety Funnel component.

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 people using Bing services, such as engaging in harmful, fraudulent, or illegal activities, violating privacy, or creating inappropriate content. Individuals 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 maintain that Bing remains a safe and transparent service for all users.

The creation experience incorporates explicit prohibitions that are not applicable in the same way to Search, including restrictions on generating images of certain celebrities and copyrighted material. When user inputs fall within restricted categories, Creator responds through refusal or equivalent non-production behavior in accordance with its creation policy.

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 support 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 offers users direct control over what they see through SafeSearch, a feature that uses machine learning to automatically detect 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 facilitate consistent application of filters. Bing also provides a mechanism for users to report inappropriate content or suggestions, including explicit content, NCII-related concerns, and other sensitive material. These reports feed into the continuous improvement of the underlying safety systems, helping Bing respond quickly to emerging issues and regulatory expectations.

During the 2026 assessment period, Bing strengthened protections for users determined to be under 18 by defaulting signed-in under-18 accounts to SafeSearch Strict where technically supported, with
region-specific implementations as appropriate. This change is intended to reduce exposure to

explicit content while preserving access to age-appropriate information through standard search functionality.

To improve clarity for users and align with the evolving scope of filtering, Bing has updated documentation terminology from "adult" to "explicit content". "Explicit content" includes, but is not limited to, adult content and more accurately captures the broader range of material subject to SafeSearch controls.

Bing is also updating protections for users in SafeSearch Strict and Moderate modes to improve consistency and mitigate unexpected exposure to harmful, offensive, or illegal content, while preserving users' ability to access sensitive content where settings permit.

***Content provenance and watermarks*:** Bing invests in helping people identify AI-generated content. Bing includes content credentials and a pixel-based watermark on generated images. The hidden watermark is imperceptible to human eyes, allowing verification of AI-generated images even after minor modifications. The watermark complements Bing's content provenance technology to help promote an AI-generated image's integrity and traceability. Bing participates in cross-industry efforts to improve identification techniques. For example, Microsoft is a founding member of the Coalition for Content Provenance and Authenticity (C2PA), an industry initiative that develops open standards to support content provenance and authenticity.

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, making sure 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 contribute to 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.

Data governance:

***GenAI disclosures*:** Copilot Search and Bing Image and Video Creator provide in-service disclosures and reminders to people that AI can make mistakes and 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 people 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 an individual 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.

***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 summarized search features 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**.** Bing has also invested in problem-specific classifiers to power defensive search interventions, such as identification of nudification tooling-related queries and results.

***Terms and 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.

***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 people using Bing services, such as engaging in harmful, fraudulent, or illegal activities, violating privacy, or creating inappropriate content. Individuals who violate these terms may face limitations on their access to the services.

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 individuals 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 make sure 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 make sure 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 seek to make sure 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 facilitates 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 a full DPIA as a 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 aims to promote 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.

***LLM-Based Guardrails and Safety Mediation*:** This mitigation implements an automated, upstream security and alignment layer powered by an LLM to evaluate user inputs against established Responsible AI (RAI) safety policies. Positioned at the pre-processing stage of the application workflow, this layer serves as a primary defense mechanism against harmful, inappropriate, or borderline policy violations before any core downstream generation or content synthesis takes place. This mitigation is intended to reduce the likelihood that generative AI experiences produce content that violates RAI policy.

Guardrails and safety mediation for Creator experiences are monitored through Bing's existing effectiveness reviews and tuning workflows. Where relevant, controls are adjusted based on observed risk patterns (e.g., during elections and other heightened-risk periods) and through continuous evaluation using classifier performance, error budgets, and operational signals to identify opportunities to further reduce residual risk while maintaining product utility.

***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 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.

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 help 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 make sure 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 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 maintain 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 help the blocklists reflect heightened sensitivities around electoral manipulation. This multidisciplinary 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 limited functionality in Bing Image and Video Creator is available to all unauthenticated users, in Copilot Search, full functionality is restricted to authenticated users who meet the appropriate age thresholds. Consistent with Bing's approach of applying the most protective standard where feasible, age-gating and associated creation constraints reflect stricter regional requirements and are deployed globally where technically and legally viable, with country-by-country tailoring as

required. These controls are aligned with Bing's strict mode of safety settings to maintain consistent protection across creation and search experiences. 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 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 users determined to be under 18 based on the birthdate in their Microsoft account, consistent with Microsoft's Privacy Statement on children and advertising. Across Bing features, authenticated users under 18 do not receive targeted advertising, and this approach is applied to new Bing features as they are introduced. Age-targeting segments for individuals determined to be under 18 are not offered within Microsoft Advertising. Under the Microsoft Advertising Remarketing policy, advertisers must not (i) create a remarketing list, retarget, or otherwise profile any individual determined to be under the age of 18; (ii) target individuals under the minimum age required for the product advertised; or (iii) implement remarketing on sites or applications directed to minors determined to be under 18 or other applicable age limitations in a given market. Microsoft also maintains clear policies prohibiting behaviorally targeted ads, or ads containing explicit material, such as alcohol or gambling, from being shown to users determined to be underage, and uses both manual and systematic methods to review ads for compliance, reserving the right to remove non-compliant content. These protections are enforced both contractually and through the absence of under-18 targeting options in the advertising manager.

***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 supports that each team member is well-versed in appropriate data use practices, reinforcing the company's commitment to safeguarding user privacy.

***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, for the insights to 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 summarized search feature(s) 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 harm 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.

In the reporting period, Bing maintained a "low leakage" posture for Creator experiences and tuned listening to capture leakage and defect trends from classifier telemetry and abuse-pattern analysis. Where patterns of misuse were observed, Creator shipped targeted classifier refinements rather than broad policy changes, reflecting the policy-first approach in this space. This included additional

and updated classifiers focused on copyright-related risks and new audio-related harms in support of video generation (e.g., leveraging transcript-based detection for audio content and separate signals for copyrighted music). These signals feed Creator's monitoring dashboards and support faster adjudication and mitigation cycles when anomalies are detected.

Overall, social listening for Creator is characterized by: (a) policy separation from Search to reflect distinct harm exposure in creation, (b) zero-tolerance presentation standards (no "off-mode" content in the creation surface), (c) classifier-led signal ingestion to detect leakage and emerging abuse vectors, and (d) iterative tuning of thresholds as model capabilities evolve. These adaptations help Creator's listening posture remain responsive to modality-specific risks while preserving consistency with Bing's broader safety governance.

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 maintain 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 promote the safety of the service.

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

Governance

***Microsoft Threat Assessment Center (MTAC)*:** 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 in both the target populations and the entities engaging in maligning influence, the team can detect new actors and methods – assessing and, in many cases, attributing influence activity. MTAC has also made additional investments in AI-based threats research that help informs certain defensive search mitigations in Bing. 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.

In 2026, Bing continued to strengthen the threat intelligence inputs that support product safety and systemic risk mitigation. In addition to existing internal threat intelligence capabilities, Bing expanded external provider partnerships to include AP, ISD, and Lead Stories, increasing coverage for fact-checking and information integrity signals. These inputs support defensive search interventions, particularly where threat intelligence identifies emerging information manipulation risks, viral social content, or election-related risk signals. This expanded threat intelligence coverage helps inform Bing's escalation, monitoring, and intervention processes within the broader product safety governance framework.

Policies and 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. Additional to our targeted algorithmic ranking interventions to reduce the risk of users inadvertently encountering misleading or harmful information on election-related manipulated topics, Bing may implement special informational and answer segments for select elevated risk elections and, depending on local investments or agreements in defined markets, to point to specific high-authority information related to the election for users, localized by language and member state.

***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 explicit 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, making sure 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 information potentially misleading users or potentially deceptive material and complements Bing's other proactive safety efforts.

In 2026, Bing continued to support user literacy through public safety messages and awareness experiences for selected high-risk topics. These experiences are intended to help users identify risk, understand available support or reporting pathways, and access authoritative resources, where appropriate.

Bing applies these mechanisms in a risk-based and market-sensitive manner. For universal or high-severity safety areas, such as suicide and child sexual exploitation, Bing seeks broad coverage across EU countries. For other topics, deployment may depend on the prevalence of the risk in a

market, local regulatory or civil society interest, and the availability of verified local partners or support resources. For example, Bing has expanded awareness of messaging for certain NCII-related queries and has implemented eating disorder public safety messaging in selected EU markets where appropriate partners were identified. These literacy-oriented interventions complement Bing's other product safety mechanisms, including authoritative information experiences, targeted public service announcements, and policy-based interventions designed to reduce exposure to harmful or misleading content.

***In product indicators – Child safety*:** For users search queries for Child Sexual Abuse Material (CSAM), Bing displays PSAs, similar to those regarding search queries with potential intent 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 show unrelated images from popular web content. Bing continues to rely on in-product policy mechanisms and safety signals to identify and mitigate child safety risks. In 2026, the DIRT team identified an additional policy framing relevant to these controls: a risk-intent model intended to support evaluation of user behavior, content, and contextual indicators when assessing potential harm. This model should be reflected as an additional mechanism supporting child safety risk mitigation, rather than as a replacement for existing child safety policies or enforcement processes.

***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 and
self-harm. Additional PSAs have been deployed in specific countries based on local interest and partnerships, such as non-consensual sexual imagery and eating disorders. PSAs are one of Bing's most visible tools for helping people 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.

In 2026, Bing expanded certain public safety messages while continuing to apply a market-by-market approach based on risk prevalence, local regulatory or civil society interest, and the availability of verified partners and support resources.

During the assessment period, Bing expanded PSA mitigations addressing eating disorders, suicide and self-harm, NCII, and other mental and physical wellbeing risks. These interventions help people identify appropriate health information, crisis-support services, reporting or removal pathways, and other authoritative resources. The applicable query sets, PSA content, and destination resources are assessed by relevant product, policy, safety, and partnership teams. Before deployment, messages and destination resources are localized, translated, and adapted to the legal, cultural, linguistic, and service context of each applicable country.

For suicide and self-harm, Bing displays PSAs at the top of search results across applicable markets, including within the EU. These notices provide authoritative information, crisis-support resources, and options for obtaining immediate assistance. Bing expanded this coverage during the assessment period, including in smaller EU markets where such notices were not previously shown. Where verified local resources are unavailable, a PSA may direct people to an appropriate international support service.

For eating disorders, Bing displays PSAs in applicable markets to direct people to authoritative health information and support resources. These notices are intended to help people access reliable guidance when searching for content that may pose risks to their mental or physical wellbeing.

Eating-disorder PSAs are currently active in Germany and France, where suitable partner resources have been identified.

For queries related to NCII, Bing displays PSAs in applicable markets to direct users to authoritative information, reporting options, victim-support services, and available removal pathways. Additional NCII notices may be deployed where local interest, risk signals, and appropriate partnerships support implementation. Within the EU, this messaging is currently available in the Netherlands through Bing's partnership with Off-Limits.

Bing also deploys safety messaging concerning child sexual exploitation and abuse material, including notices explaining the illegal nature of such material and directing users to relevant support resources. Bing has partnered with the Lucy Faithfull Foundation to support an international version of applicable safety messaging.

Because these interventions depend on verified content, suitable support partners, and the availability of appropriate destination resources, coverage varies by market and risk area. Bing prioritizes broader deployment for acute and widespread harms, such as suicide and child sexual exploitation, while applying more targeted implementation for other risks based on local demand, risk signals, and partner availability.

External partnerships and wider information-safety measures

Bing complements its product-level interventions with a multi-stakeholder approach involving technology companies, governments, civil-society organizations, researchers, and subject matter experts. These external relationships provide threat intelligence, specialist knowledge, incident-response mechanisms, and authoritative resources that can inform Bing's assessment and mitigation of risks related to public security, violent extremism, and information integrity.

As a member of the Global Internet Forum to Counter Terrorism (GIFCT), Bing participates in an industry initiative that brings together technology companies, governments, and experts to prevent terrorists and violent extremists from exploiting digital platforms. Through this membership, Bing has access to GIFCT incident-response mechanisms, including processes for receiving hashes associated with events activated under applicable content-incident protocols. These capabilities enable Bing to become aware of relevant material more quickly, assess whether it may affect the search experience, and take appropriate action in response to content circulating online following a terrorist or violent-extremist event.

Bing also participates in the Advisory Network of the Freedom Online Coalition, a multi-stakeholder advisory group supporting a coalition of governments working to advance internet freedom. In addition, Bing works with the Institute for Strategic Dialogue (ISD) to improve its understanding of terrorism and violent-extremism-related risks in search. These engagements can help Bing identify evolving narratives, assess potential search-related risks, and support counter-narrative initiatives, including through advertising grants that surface credible alternative information for users whose queries indicate interest in extremist organizations.

These partnerships reinforce Bing's broader approach to public security by combining internal monitoring and incident-response capabilities with external expertise, shared intelligence, and coordinated industry processes. They help Bing respond to fast-moving events while maintaining a proportionate approach that takes account of user safety, freedom of expression, and access to information.

Microsoft also supports reputable third-party assessments of the credibility of online news sources. Within Bing, trustworthiness and authority indicators form part of the signals used to help users evaluate the reliability of news and information encountered through search. Together with ranking systems, threat intelligence, provenance tools, transparency measures, and targeted interventions,

these indicators support a more informed search experience and help reduce the prominence of misleading or low-authority information.

Microsoft Advertising policies:

The Microsoft Advertising policies set out the requirements for ad content, including the criteria for content removal, and require advertisers and partners to comply with these policies throughout their use of the services. Microsoft Advertising maintains clear and regularly enforced content policies and practices to prohibit and mitigate advertisements that harm users or violate content standards.

Microsoft Advertising also maintains relevance and quality policies governing the relevance and quality of advertisements served through its network. The policy framework addresses financial scams through three reinforcing policy areas:

The finance policy: Disallowing get-rich-quick schemes, Ponzi and pyramid schemes, and multi-level-marketing models, employment or income opportunities requiring up-front payment, and services guaranteeing financial returns or foreclosure prevention.

The information integrity and misleading content policy: Prohibiting unsubstantiated claims, fraudulent "free" offers or misleading pricing claims, misrepresentations, misleading ads, and disinformation landing pages.

The malware, malicious and unwanted software, and phishing policy: Prohibiting credential-phishing sites, tools, and related behavior.

Terms of use and content policies are published online, readily accessible to advertisers and Bing users, updated as needed, and translated into multiple languages. Advertisers retain ownership of and responsibility for their ad content and must agree to the applicable terms when signing up for a Microsoft Advertising account. Microsoft Advertising employs dedicated operational support and engineering resources to enforce these policies through a combination of automated and manual methods, and updates them over time, based on trend monitoring, detected high-severity-harm signals, and engagement with external partners and regulatory bodies.

For the protection of minors, Microsoft does not deliver personalized advertising to users under 18 whose birth date in their Microsoft account indicates they are minors. Advertisers are prohibited from creating remarketing lists targeting individuals under 18, from targeting individuals below the minimum age for a given product, and from implementing remarketing sites or apps directed to minors. Microsoft applies both manual and systematic review to remove non-compliant ads and to prevent behaviorally targeted or explicit-material ads (e.g., alcohol, gambling) from being shown to known child users. These protections apply across Bing features and will continue as new experiences are introduced.

For political advertising, Microsoft does not allow political advertising within the Microsoft Advertising ecosystem (including Bing). Ads for political candidates, parties, ballot measures, political fundraising, and election-related content are prohibited globally, alongside certain issue-based advertisements. Microsoft Advertising's relevance and quality policies also help prevent misleading, deceptive, or low-quality advertising, and drive continuous improvements through new signals, rules, and model updates.

***Microsoft Advertising restrictions on political ads*:** Microsoft does not allow political advertising within the Microsoft Advertising ecosystem, which supports all advertising on Bing. Microsoft Advertising policies prohibit ads for election-related content, including political candidates, parties, ballot measures, and political fundraising globally. Ads supporting fundraising for political candidates, parties, political action committees, or ballot measures are also prohibited. Microsoft Advertising policies also prohibit
issue-based advertisements, as defined in its published policy. Microsoft Advertising is a signatory to

the EU Code of Conduct on Disinformation (COCD) under the DSA , and further details on its commitments are set out in Microsoft's most recent COCD report.

Over the past year, Microsoft Advertising maintained this prohibition and introduced additional compliance measures in alignment with EU Regulation (EU) 2024/900 on the Transparency and Targeting of Political Advertising (TTPA), which entered full application on 10 October 2025. The definition of political advertising applied to ads serving the European Union was updated to align with the scope set out in the regulation.

To enforce this prohibition at an earlier stage in the campaign life cycle, a mandatory advertiser Political Advertising Declaration was introduced in September 2025. Advertisers must declare, before campaign creation or import, whether a campaign contains political advertising. Because Microsoft Advertising does not permit political advertising, any campaign declared as political is blocked from serving. This establishes a proactive, declaration-based control at the point of campaign entry, rather than relying solely on downstream detection. Non-compliance, including misrepresentation of a campaign's political nature, can result in account suspension or termination. Campaigns active prior to these changes remain subject to ongoing review to confirm continued compliance. These measures are reflected in updated policy documentation.

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 making sure 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. **LLM-based guardrails and safety mediation:** This mitigation implements an automated, upstream security and alignment layer — powered by an LLM — to evaluate user inputs against established Responsible AI (RAI) safety policies. Positioned at the pre-processing stage of the application workflow, this layer serves as a primary defense mechanism against harmful, inappropriate, or borderline policy violations before any core downstream generation or content synthesis takes place. This mitigation is intended to reduce the likelihood that generative AI experiences produce content that violates RAI policy. ***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 people 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 individual 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 people are engaging with AI-generated content. For example, in Copilot Search, an information icon near the feedback buttons alerts people 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 promote accessibility 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 people 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 life cycle (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 people to trusted sources through several tools. For high-risk events, such as various high-risk EU elections, Bing highlights answers and PSAs that point to expert-vetted information. During events like public health crises or violent conflicts, Bing would apply defensive search interventions, such as demoting low-authority results in ranking on queries that may lead people to data voids and misleading content informed by threat intelligence.

Bing continues to prioritize credible, authoritative, and relevant content as a core product safety mitigation. Bing's approach is designed to balance users' rights to access information and expression with the need to reduce exposure to harmful, misleading, or unexpectedly offensive content.

For generative AI experiences, Bing mitigates risk by grounding AI-generated responses in high-authority, trustworthy third-party content, where applicable. Grounding supports accuracy, relevance, and user transparency by tying generated responses to source material and enabling people to review cited sources.

***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 OKRs for the year in addition to dedicated work cycles within to track and move progress, 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 facilitate appropriate allocations of resources, especially key areas of content concern. Finally, Bing sends externally-facing daily updates on key areas of content and conduct concern to support appropriate prioritization, particularly improvements to GenAI features.

In 2026, the DIRT team reaffirmed the strategic importance of regularly reviewing and updating product safety metric targets to reflect evolving regulatory standards, new risk signals, and stakeholder expectations. Additional emphasis was placed on working with cross-functional partners to help metric thresholds account for shifts in external risk indicators, including emerging trends in digital harm and new compliance obligations discussed during recent DIRT and status meetings. While the prior metric definition cycle was maintained, the group agreed that each review must now explicitly capture inputs from legal/regulatory intelligence and integrate feedback from the Digital Safety Unit and cross-market audits. New for this cycle is an open action to synthesize ecosystem map metrics into a narrative for the SRA and document year-over-year changes and investments in the executive summary. Deprecation of obsolete metrics and products — as flagged in weekly compliance meetings — should be handled transparently, with zeroing out in the audit matrix and removal from risk registers and manifestations. The responsible metric owner is required to escalate unresolved metric requests, coordinate biweekly executive updates, and track all modifications in the mitigation matrix for full traceability.

***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 teaming and systematic sweeps are conducted to identify and address defects, particularly in advance of high-risk situations, such as elections or major breaking news events. For example, during initial weeks of Iran war as part of defensive playbook for elections and crisis events, and emerging problem areas, such as nudification tooling in search.

In 2026, Bing continued to mature its product safety testing and monitoring approach by applying a more risk-driven model to emerging content and search safety issues. The team described expanded use of operational moderation, red team-style review, and event-driven sweeps to identify and triage search risks during periods of elevated risk, including emerging attack patterns, crisis events, and high-profile elections.

These activities are supported by improved processes for monitoring user feedback and external signals, including social media feedback, threat intelligence inputs, and event-specific query sets. The team has refined these processes to improve signal-to-noise ratio, prioritize actionable cases, and route findings to product teams for mitigation, where appropriate.

Red team findings are used to refine training data, harden safeguards, and improve system behavior. Bing maintains 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 makes sure 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 make sure 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 make sure that new features, mitigations, and metrics are evaluated routinely and are functioning as designed.

In 2026, Bing continued to mature this evaluation approach through a policy framework that considers both risk to users and user intent. This risk-intent model supports more nuanced evaluation of safety performance by helping teams distinguish between user-directed access to information and scenarios where users may be unexpectedly exposed to harmful, misleading, offensive, or otherwise unsafe content.

Bing also continued to strengthen the connection between policy and measurement. The DIRT team has developed a manually curated, policy-aligned benchmark query set that includes real and hypothetical query traffic and is used to support ongoing monitoring of safety stack performance. This benchmark is intended to help Bing evaluate whether product mitigations remain aligned with current policy expectations and to support updates as policy evolves.

For generative AI experiences, Bing is also developing grounding metrics to evaluate how well AI-generated responses are grounded in high-authority, trustworthy third-party content. This work is not yet fully operationalized, but the design specification has been completed and is expected to support future evaluation of generative AI safety performance

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 classifiers additional to threat intelligence, 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.

Bing continues to use defensive search interventions as targeted mitigations within the content and activity moderation category. In 2026, Bing further refined these interventions through a policy framework that evaluates both the risk presented by content and the likely user intent associated with the query. This risk-intent approach supports more tailored mitigation decisions, including when to preserve access to information and when to apply interventions to reduce potential harm.

Bing also expanded public safety messaging as part of its targeted intervention tool kit. These interventions include public safety or awareness messages for certain high-risk query categories, including
NCII, suicide, eating disorders, and child sexual exploitation-related harms. In addition, Bing also expanded EU coverage for suicide-related public safety messages, eating disorder public safety messages active in Germany and France, and NCII-related awareness messaging in the Netherlands through a partner-supported deployment.

Safeguards – *Shopping*: Shopping uses a two-filtering system where ads team conduct a first vetting and Bing conducts a second vetting process (97% content comes from ads and 3% 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. Flagged viral potential issues are manually reviewed, analyzed daily, and shared across Bing's product teams and product engineering leadership in a daily report. These reports inform Bing of the public concerns, issues, or emerging trends and can serve as a 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 events 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.

Microsoft continued to strengthen its threat intelligence capabilities used to inform defensive search interventions for information operations and information manipulation risks. In 2026, the team expanded the scale of threat intelligence inputs used to identify fast-emerging manipulated narratives, including viral and socially amplified topics, and to translate those signals into defensive search interventions on related queries.

Microsoft also expanded its external threat intelligence and fact-checking coverage through a broader suite of provider partnerships. These partnerships support improved visibility into manipulated narratives, information integrity risks, and emerging harm areas, while also reflecting continued investment in the fact-checking and trust and safety ecosystem. Partner-specific names should be included only where Microsoft has confirmed that the partner has approved the proposed characterization.

In addition, Microsoft strengthened engagement with regulators, digital services coordinators, civil society organizations, academic institutions, and other trust and safety ecosystem participants. These engagements provide external signals that help inform product, policy, metrics, and intervention decisions. The team also uses external speaker series and ecosystem networks to gather insights on emerging risks, including information manipulation, election-related risks, fraud, NCII-related tooling, suicide-related harms, and other areas where external stakeholders identify evolving risk patterns.

Governance

***Threat intelligence*:** Bing maintains a dedicated threat intelligence capability to identify and respond when high-risk search topics, emerging events, or coordinated manipulation may cause search results to fall short of Bing's Trustworthy Search Principles. The team focuses particularly on circumstances in which malicious actors seek to exploit data voids, manipulate ranking signals, or elevate low-authority, harmful, or misleading content. Its objective is to reduce the risk that users are unexpectedly exposed to such material, while preserving access to lawful information that users intentionally seek.

The team combines proprietary detection pipelines, targeted monitoring, internal intelligence, and signals from trusted external organizations to identify emerging threats and assess their potential relevance to search. These capabilities support the detection of viral narratives associated with information manipulation, coordinated influence activity, cyber-enabled fraud, hate and extremism, and crisis events that may create heightened search risks. Microsoft also works to identify and track nation-state information operations targeting democratic institutions and processes around the world. Intelligence from this work, together with early-warning indicators supplied by trusted

partners, can inform Bing's detection priorities, defensive-search strategies, and operational readiness.

When a validated risk is identified, the threat intelligence team can deploy tailored algorithmic interventions. Depending on the nature of the threat and the apparent user intent, these measures may include boosting authoritative sources, demoting low-reliability or manipulated content, restricting problematic search suggestions, or applying alternative ranking treatments. Interventions are intended to prevent users from encountering harmful or misleading material unexpectedly, rather than to broadly limit access to information. The team monitors the performance of these measures and refines them as the underlying event, narrative, or abuse technique evolves.

Bing has continued to scale these capabilities through expanded relationships with fact-checking organizations, academic and independent researchers, threat-intelligence providers, and other trusted experts. These collaborations provide additional signals concerning emerging misinformation campaigns, coordinated manipulation, new abuse techniques, and rapidly spreading narratives. Bing assesses these external inputs alongside proprietary detection systems, user reports, social-listening insights, red-team findings, and operational metrics to improve the speed, precision, and coverage of its response.

The threat intelligence team also works closely with internal Microsoft partners. Collaboration with the Microsoft Threat Analysis Center (MTAC) supports monitoring of geopolitical developments, nation-state influence operations, and other emerging information threats. Coordination with Microsoft cybercrime and cyber-defense teams provides intelligence concerning cyber-enabled fraud, scams, malicious infrastructure, and other activities that may affect search users. These partnerships help Bing translate broader threat intelligence into targeted search interventions, particularly for elections and civic events, fraud, geopolitical crises, and other fast-moving or high-impact topics.

During major high-risk elections, Bing enhances its threat-intelligence operations through dedicated monitoring of election-related queries, narratives, and potential information-manipulation campaigns. The team draws on proprietary detection pipelines, Microsoft intelligence, trusted external sources, and targeted monitoring to identify coordinated attempts to influence the search environment. Where appropriate, Bing can elevate authoritative election information, reduce the visibility of low-reliability or manipulated content, restrict problematic suggestions, and rapidly deploy or adjust interventions as risks emerge throughout the election cycle.

Together, these capabilities provide Bing with an integrated process for identifying, assessing, escalating, and mitigating dynamic search risks. The combination of dedicated expertise, proprietary detection, crisis-response procedures, internal Microsoft intelligence, external partnerships, and ongoing measurement enables Bing to respond with greater speed and precision while maintaining a proportionate approach that balances user safety with freedom of expression and access to information.

Policies and 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 makes sure 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.

***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 summarized search features 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. Bing has also invested in problem-specific classifiers to power defensive search interventions, such as identification of nudification tooling-related queries and results.

Responsible product/service design

***Limit lengths for generative search queries*:** Copilot Search is limited by the 2,000-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*:** Bing has a safety classifier in place that can identify elections-related queries to support targeted search feature mitigations. This classifier has been a focus of continued improvement and supports multiple languages. Bing has also invested in a safe routing and response generation system for summarized answers with grounding, citation, and response guardrails on high-risk queries.

Bing's targeted election interventions continue to be supported by threat intelligence and defensive search workflows. These workflows include monitoring emerging or manipulated narratives, translating relevant intelligence into product-facing insights, and applying algorithmic or operational interventions, where appropriate. Bing also participates in election rapid-response channels,

including established DSA-related channels, and reported receiving almost no authority notices related to search incidents during elections through these processes.

For GenAI election experiences, Bing is continuing to work to support safe-election-related summarization, including grounding and source-quality controls. Broader use of summarized responses for election topics should be described as a future-state objective and ongoing mitigation development effort, not as a control that is currently in production or broadly available.

***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 make sure 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 to evolve mitigation practices to meet the rapidly adapting tactics to circumvent Bing safeguards and access potentially harmful content. Bing also has investments in viral and emerging events-based detection of threats.

In 2026, Bing continued to mature its approach to monitoring emerging threats by combining policy-aligned measurement, external threat intelligence, and event-driven operational response. The DIRT team described expanded use of curated benchmark query sets and manually labeled examples to assess how interventions perform against evolving policy expectations. These measurement approaches are intended to support more responsive monitoring as policies and threat patterns change.

Bing's emerging-threat monitoring also incorporates crisis-event playbooks and event-driven moderation sweeps. These processes are used to identify risk spikes, assess defensive search exposure, relay findings to relevant product teams, and inform threat-intelligence briefs or other operational interventions. Recent crisis-event activity was used to test and refine this approach.

The DIRT team also noted increasing use of AI-assisted capabilities and agentic tooling to improve threat-intelligence processing and operational efficiency. This should be framed as an enhancement to risk management capability, rather than as a separate product-risk claim, unless specific tooling and use cases are confirmed.

***Content moderation – Advertising – Continuous improvement*:** Microsoft Advertising continually adds new signals and information sources to improve its detection systems, evaluates the incremental performance of each candidate signal and information source against existing proactive, targeted reviews for emerging risks and detection models, and adopts those signals and information sources that strengthen the system.

This investment has expanded from signal evaluation to the deployment of new detection capabilities (including LLM-based detection of phishing, spoofing, and related fraud patterns); mandated verification of advertiser accounts; spend-velocity checks for anomalous budget spikes;

and detection of redirect and landing-page mismatches that indicate misleading behavior. Model development is governed by a structured release process in which candidate models undergo a testing period for human evaluation and bias detection before live deployment, with high-precision confidence thresholds established through testing on curated datasets.

Detection services are continuously monitored, retrained on recent fraud data, and supplemented by re-review of historical content when new risks are identified, ensuring consistent enforcement over time. LLM labeling guidelines serve as the foundational model training for automated enforcement decisions and are updated on an ongoing basis to reflect evolving fraud tactics, informed by engagement with external partners and regulatory bodies.

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 – Content removal policies and procedures*:** 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, to prevent such content existing in Bing's index and from appearing in the search results. This scanning occurs before content enters the search index and is used to make sure 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 third-party partners experts, 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 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. To support these efforts, Bing partners with StopNCII and uses matching technologies to detect content contained in the StopNCII database, helping prevent known NCII from appearing in search results. Bing also maintains processes for victims and third parties to report NCII for review and removal.

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.

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 people of certain risks through PSAs or warnings, or provide people with options for tailoring their content.

***Report a concern*:** Bing provides people with an accessible report about a concern mechanism on every page of Bing Search and across GenAI experiences. Through this function, people 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: Microsoft Advertising reserves the right to issue immediate suspensions of an advertiser's account where the advertiser egregiously violates Microsoft's terms or policies, including where those violations indicate intent to use Microsoft services in a fraudulent, harmful, or illegal manner. Microsoft uses both manual and automatic methods to review ads for compliance with its policies and regularly removes non-compliant content. Account-level enforcement operates through a defined model: (1) ad and ad-component disapprovals (which identify the specific policy violated and provide an appeal path), (2) repeated-violation "three strikes" (a first strike suspends all ads and ad components within the impacted advertiser account until remedied, a second strike includes the previous actions and prevents the Manager Account from creating new accounts preventing customer level re-offending or cross-account non-compliance, and a third strike in any policy category results in immediate suspension of all of the advertiser's accounts without further warning without access to any further ad delivery options), and (3) immediate account suspension for egregious violations that pose an immediate threat to the safety or security of users, Microsoft, or its partners. Financial-scam violations are treated as egregious and result in immediate suspension. Enforcement actions are executed and tracked within our internal enforcement and compliance tools and are governed by standard operating procedures that support consistent application and structured handling of advertiser appeals.

Bing's enforcement posture for advertising continues to prioritize immediate account suspension in egregious cases of fraudulent, harmful, or illegal conduct, supported by both automated and manual review. During the period, regulatory scrutiny of fraud in search and advertising increased, including formal information requests to a limited set of companies. In response, Bing reinforced its expectation that advertisers adhere to Microsoft Advertising policies and expanded internal emphasis on demonstrating the effectiveness of mitigations and the methodological rigor of enforcement evaluations.

While Microsoft Advertising is a distinct organizational team, Bing is the regulated service for DSA purposes. Accordingly, Bing treats policy-violating advertising content served on Bing as in scope for systemic risk mitigation and coordinates with Microsoft Advertising and Microsoft Merchant Center to ensure policy compliance, with immediate suspension in egregious cases and removal of non-compliant ads. This includes intensified attention to scam patterns, such as spoofing and impersonation, that can drive consumer losses and bank chargebacks.

Bing is enhancing its assessment lens on consumer harm for advertising-related fraud, recognizing that some outcomes occur off-platform and require cross-team collaboration to detect and measure. As part of this work, Bing is engaging with Microsoft Advertising to improve visibility into user-harm signals and to align enforcement posture, appeals, and escalation pathways for high-severity fraud scenarios, while preserving privacy and complying with legal and competition constraints. Bing will update this section as these enhancements to metrics and workflows are operationalized.

***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 makes

sure 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 of 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 summarized search feature(s) 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.
- Social listening findings are routed through established incident response and decision pathways (e.g., Bing's Safety Playbook) to enable timely triage, prioritization, and enforcement when trends indicate elevated risk.
- Where appropriate, AI-assisted triage and signal enrichment are used to improve threat intelligence processing and analysis throughput, as an enhancement to risk management operations rather than a separate product risk claim.
- Validated social listening findings can be actioned via Bing's Content Moderation Platform (CMP), enabling policy consistent implementation of URL/query actions and related enforcement updates across Core Search and GenAI surfaces.
- Social listening complements red teaming and other intake channels as part of Bing's ongoing monitoring of metrics, incidents, and potential risks, with mitigations tuned to work across languages
  and markets.

***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 people of certain risks through PSAs or warnings, or provide people with options for tailoring their content.

Bing continues to maintain content moderation policies and procedures designed to support consistent review, escalation, and removal of decisions for content that may violate applicable policies or legal requirements.

In 2026, Bing's content moderation approach should also reflect the DIRT team discussion of an updated policy framework organized around a risk-intent model. This framing supports a more structured assessment of whether content presents a relevant risk, the likely intent or use case associated with that content, and the appropriate mitigation response. The framework is intended to help align content removal decisions with Bing's broader user safety commitments and DSA-related systemic risk mitigation obligations, while preserving procedural consistency across review, escalation, and enforcement activities.

Where content is identified for potential removal or restriction, Bing should continue to apply documented review procedures, including escalation pathways for higher-risk or legally sensitive content categories. The updated framing should be reflected in the SRA as an enhancement to existing content moderation governance rather than a replacement of Bing's established removal controls, unless confirmed otherwise by the policy owner.

***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 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 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: Our proactive detection scope for advertising explicitly treats ad scams, misleading promotions, and related deceptive practices as within the Bing SRA remit, with joint accountability between Bing and ads for risk identification, measurement, and response.

The Microsoft Advertising detection system is multi-tiered, combining machine-learning (ML) models, heuristic assessments and rules, and manual review to prevent and detect fraudulent ads from entering the system, with models and rules that apply both at the advertiser level and to individual ads.

Ad-level models are applied based on the Microsoft Advertising policies relevant to the market where the ad serves and include, among others, CSEAI, offensive/hate speech, violence, suffering, gore, exploitation, explicit content, gambling, tobacco, weapons, and prohibited and restricted healthcare products.

Advertiser-level controls and models include: Advertiser Identity Verification (AIV): Mandated for advertising customer accounts, which requires advertisers to prove their legal identity before running ads and is reinforced with domain-verification checks (including the Domain Verification Tool), which confirm domain ownership and registration. These controls deter phishing, personal-information fraud, and disallowed-product advertising, payment fraud detection.

Purchase risk model: Triggered at billing to determine whether to attempt the charge or block the advertiser.

Behavioral risk model: Evaluating ad text, keywords, budgets, account age and structure, payment-instrument attributes, associated individuals, and access devices whenever an advertiser changes campaigns.

LLM-based spoofing detection: Targeting advertisers impersonating legitimate sites.

LLM-based celebrity-scam and brand-impersonation detection: Multilingual models that identify false celebrity endorsements, brand and company impersonation and phishing, and counterfeit patterns, reducing scam impressions.

Account-compromise (account-takeover) detection: This was introduced in response to a rise in account compromises to overall compromise impact, and serves as an autonomous AI review agent that supports domain review processes and enables automated enforcement actions in defined scenarios.

Risk detection is organized through a structured fraud taxonomy of "constellations" (groupings of related fraud patterns used to index, escalate, and report complex scenarios) and "patterns" (granular tactics,

such as cloaking, phishing, and baiting), supported by an internal tracker that captures suspicious accounts not matching known patterns.

The classifiers and rule-based systems evaluate:

- Behavioral signals – Spend rate, budget/bid patterns
- Identity signals – Account, customer, payment instrument and device
- Text and visual content – Ad text, images, landing-page screenshots

Link signals – Redirects, cloaking, lookalike domains, with heightened scrutiny of high-severity-harm traits, such as guaranteed return claims, violations of regulatory requirements, such as disclaimers or risk warnings, urgent financial calls to action, newly created or rapidly scaled accounts, repeated violations, payment anomalies, and cloaking.

Based on model output, Microsoft Advertising may take enforcement action, including account suspension. The system also incorporates internal and external signals. Internal signals include a Bing signal indicating whether a domain carries spam and a Microsoft Defender signal indicating user harm from visiting a page and external signals include aggregated malware-risk signals from VirusTotal, domain-risk tools, and domain age/expiry and cross-engine click data from Whois.

***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

Bing applies targeted moderation to advertising surfaces to prevent misuse that could harm users or compromise the integrity of the service. Consistent with Microsoft's policies, advertisers who engage in fraudulent, misleading, or illegal behavior are subject to immediate suspension of advertiser accounts, with enforcement decisions made under documented guidelines and appropriate oversight to ensure consistency and due diligence.

Regulatory focus on fraud, dark patterns, and consumer protection continues to increase and is expressly expected to span both search and advertising flows. Authorities expect stronger evidence of consumer-harm measurement and view the designated service as accountable for user outcomes irrespective of internal organizational boundaries. In response, Bing is strengthening its measurement and reporting to provide clearer linkage between mitigations and observed user-harm outcomes, including within advertising-related pathways.

Recognizing the critical importance of maintaining a trustworthy ad ecosystem, Microsoft Advertising enforces stringent content policies that prohibit misleading, deceptive, and fraudulent advertisement content. These are supported by Relevance and Quality Policies and applied through structured review processes that assess whether ads meet legal, relevance, and quality standards, and that detect and prevent misleading or manipulative practices. Microsoft Advertising employs a combination of human and automated reviews using algorithmic systems and a robust filtration system to detect and mitigate harmful behavior such as fraud, phishing, malware, or compromised accounts. Bing users can report advertising for takedown through the "feedback" function. Detection of violative content covers ad text, images, extensions, product feeds and components, landing

pages, URLs, keywords, and targeting settings. Malware is treated as an egregious violation and results in suspension of the advertising account.

Moderation now operates as a layered pre-publication, post-publication, and reactive enforcement processes in which human experts design, test, and operate automated systems, validate flagged content, and feed results back to improve model performance, detect drift, and guide retraining, while supporting governance, compliance, and incident response. Manual review is engaged where automated systems lack sufficient confidence to take enforcement action and in response to advertisers' appeals, with reviewers working from standardized operating procedures. LLM labeling guidelines are used to create curated reference datasets allowing for rapid propagation of newly identified patterns for classifier training to support consistent decisions and to create LLM prompts used to detect missed non-compliant content. Review coverage extends across multiple escalation channels, including user report web forms, government inquiries, and internal escalations, complementing automated detection with human oversight of complex or borderline cases.

***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 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.

Bing maintains Family Safety settings as a core mitigation within Incident Response and Content and Activity Moderation, with a focus on age-appropriate experiences and protections for minors across Bing surfaces. Bing is implementing a default SafeSearch Strict setting for users under the age of 18 in signed experiences. This change strengthens baseline protections for minors and aligns with the broader regulatory emphasis on child safety and age-appropriate experiences.

Bing recognizes that child safety expectations are evolving globally, and that regulators increasingly benchmark search services against child safety guidance when assessing systemic risk mitigations. Consistent with this environment, Bing's Family Safety settings:

- Default to SafeSearch Strict for under-18 signed-in accounts and uphold this enforcement across surfaces where technically supported.
- Are informed by cross-jurisdictional privacy and competition constraints, including the use of age signals within Microsoft account experiences in a manner that respects applicable data minimization, privacy, and competition requirements.
- Are part of an active, ongoing program of experimentation and learning to determine privacy-protective, effective approaches to age assurance and youth protections in search, taking into account differing regional legal frameworks and the current lack of a single accepted global standard.

To preserve compliance boundaries, Bing does not rely on identity signals from services with separate account systems, and any use of account level-age indicators is limited to appropriate signed-in contexts and subject to applicable regulatory frameworks. Bing continues to evaluate these controls and their effectiveness, including the clarity of service safety settings, with the objective of reducing minors' exposure to harmful content and supporting parents and guardians in managing youth experiences.

Bing's Family Safety settings are reinforced with operational monitoring and response processes. These include safeguards for minors' accounts through default SafeSearch Strict and escalation pathways within incident response for content that presents heightened child safety or legality concerns, consistent with applicable law and policy.

Incident Response and Service Level Agreements (SLAs): At the core of Incident Response and 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 (OKRs).

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 people who use Bing 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.

In Microsoft Advertising, account-level enforcement operates through a structured model: (1) ad and ad component disapprovals with notice and an appeal path; (2) a repeated-violation three-strikes regime in which a first strike suspends all ads and ad components within the impacted account until remedied, a second strike extends the suspension and prevents the Manager Account from creating new accounts, and a third strike in any policy category results in immediate suspension of all of the advertiser's accounts without further warning; and (3) immediate account suspension for egregious violations that pose an immediate threat to user or platform safety (for example, financial scam activity). Enforcement actions are executed and tracked in internal tooling under standard operating procedures to support consistent decisions and structured handling of appeals.

***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 to 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. Users who use Bing 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 to 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, civil society organizations, academic researchers, and threat intelligence providers, to inform defensive search mitigations and broader systemic risk management.

Collaboration with third parties and adherence to external evaluations further reinforces 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 industry initiatives such as the Digital Trust & Safety Partnership (DTSP), which facilitates cross-industry collaboration on trust and safety practices, and the GNI, which audits Bing, and undergoing independent assessments by organizations such as Ranking Digital Rights, Bing subjects its practices to scrutiny, supporting alignment with global standards for Freedom of Expression and Information.

Bing maintains structured external engagements to surface emerging harms and inform moderation priorities. These engagements include ongoing dialogues with Digital Services Coordinators in EU Member States and periodic touchpoints with election authorities around national and EU ballots. Pre-election and post-election reviews requested by Member States continue to occur on a routine basis and are used to identify manipulation risks, misinformation patterns, and other civic process threats that warrant proactive intervention in search results and related experiences.

Engagements with regulators and civil society have consistently highlighted several moderation-relevant themes for early threat identification. These include: (i) expectations for child safety and age-appropriate experiences and the growing use of age assurance signals as a baseline for systemic risk mitigation; (ii) clarity and conspicuousness of illegal content reporting pathways; (iii) persistent risks related to NCII and challenges distinguishing lawful explicit content from illegal or harmful material; and (iv) elevated scrutiny of AI-enabled abuse scenarios (for example, chatbot harms and generative content misuse). Insights from these partners inform the prioritization of detection signals, reviewer guidance, and enforcement of playbooks across moderation workflows.

Bing also continuously monitors enforcement actions and investigations involving other large online services to anticipate shifts in supervisory focus (for example, child safety, deceptive design patterns, and AI safety). Lessons learned from these proceedings are incorporated into the threat of

identification criteria used by policy, safety, and moderation teams to refine classifiers, escalate edge cases, and adjust interventions where appropriate.

To strengthen protections for minors, Bing is aligning with company-level initiatives that aim to enable responsible privacy-preserving age signal use across Microsoft products. While service-to-service data sharing remains constrained by applicable privacy and competition laws, Bing is preparing to consume permitted and reliable signals for authenticated users, where appropriate, to route stronger safeguards and age-appropriate experiences into moderation and enforcement flows.

External Engagement

***Microsoft-wide Election Support*:** An important component of the dedicated cross-functional team's strategy focused on democracy and elections 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:

Campaign Success Team: Microsoft helps political campaigns navigate cybersecurity challenges and the new world of AI by deploying a new "Campaign Success Team", which is part of Bing's Election Support & Response (ESR) efforts. 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 them with Microsoft's Deceptive AI Reporting Tool (discussed below).

Deceptive AI Election Reporting Tool: Microsoft has a 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 in the detection of harmful deepfakes. This web page is available, localized throughout the EU.

Cybersecurity and Cyber-influence: Microsoft identifies, tracks, and reports on potential information operations and foreign malign influence operations working with partners.

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 election-related customers to provide security guidance and tools to improve their cyber-resilience and protect the integrity of the electoral process.

Microsoft supports the cybersecurity maturity and resilience of democratic stakeholders through nonprofit partnerships. Bing's Election Support & Readiness (ESR) team works with IFES, International IDEA, UNDP (UN Development Program), NDI (National Democratic Institute), IRI (International Republican Institute), Freedom House, Witness, and others to provide programs and trainings in service of journalists, election officials, political parties, observers, human rights defenders, and more.

Bing further strengthened its election response capability through a problem-focused operational approach. The team can now stand-up topic-based query sets for event- or election-related monitoring within approximately one day, enabling faster measurement of potential defects and more targeted defensive search interventions during high-risk election periods. These capabilities are supported by threat intelligence, classifier-based interventions, and operational review processes designed to identify and address election-related search risks.

Bing's engagement with regulators, Digital Service Coordinators, civil society organizations, and academic institutions also informs election-related risk management. These engagements help Microsoft explain how search works, understand external concerns, and incorporate relevant signals into product, policy, and measurement work. Bing received almost no authority notices related to search incidents during elections through these processes, reflecting the maturing of its election-related mitigations.

***Researcher and Academic Support*:** Bing's approach to Researcher and Academic Support which is 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 through 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 mitigations library.

***Microsoft Research*:** Bing and Microsoft, including through its research arm Microsoft Research, have collaborated with 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. Bing also works with external organizations on in-depth, topic-specific reports, red-teaming exercises, and expert consultations that provide insights and recommendations to inform product development and policy shaping. These engagements can inform the processes of identifying, assessing, and mitigating risks across Bing and provide a 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, nonprofits, 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, and Microsoft routinely receives insights and information from other industry participants, as well as briefings from researchers. 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 engages with regulators, election officials, and other public authorities to understand emerging concerns, share information, and incorporate feedback into product design, safety systems, risk assessment methodology, and mitigation evaluation where appropriate. These engagements include topics such as SRA methodology, measurement of mitigation effectiveness, elections, child safety, age assurance, illegal content, NCII, AI-related risks, and fraud. Bing monitors regulatory developments across the EU and other jurisdictions to identify areas where expectations are evolving and to ensure that Bing's risk assessment and mitigation approach remains responsive to those expectations.

Between August 2025 and July 2026, Bing continued to monitor regulatory developments across the EU and other jurisdictions to identify areas where expectations are evolving and to ensure that Bing's risk assessment and mitigation approach remains responsive to those expectations. External stakeholder feedback remains an important input into Bing's SRA process by helping identify emerging risks, validate internal risk themes, and inform prioritization of mitigation work. Bing also continued to use ecosystem input from civil society, academic stakeholders, industry networks, and external threat intelligence partnerships to support product, policy, measurement, and defensive search interventions, including interventions related to information manipulation and rapidly emerging viral narratives. The availability of local partners' verified content, and market-specific expertise may affect the pace and scope of localized interventions.

***Researcher Program to enable EU Researchers*:** Bing operates a qualified researcher program to enable EU researchers to easily access Bing datasets under Articles 40.4 and 40.12 of the Digital Services Act (DSA). Researchers can request access to Bing data that is publicly accessible on Bing, upon application, from a singular landing page. In addition, researchers can request access to datasets that are not publicly available through the DSA Qualified Research program.

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, 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.

In 2026, Bing and its cross-functional partners continued to strengthen collaboration across the broader trust and safety ecosystem, including engagement with Digital Services Coordinators, civil society organizations, academic institutions, and other external experts. These engagements provide additional inputs into Bing's governance, policy, and risk mitigation processes by surfacing external perspectives on emerging harm areas, regulatory expectations, and operational signals. Bing has used these inputs to inform product and policy considerations in areas such as fraud, NCII, elections, child safety, suicide-related harms, and access to harmful tools or intermediary services.

In addition, Microsoft's participation in trust and safety networks, including the Tech Coalition, supports shared learning and information exchange with other technology providers in relevant harm areas. These partnerships complement Bing's internal expert-team governance model by helping cross-functional teams identify emerging risks, evaluate potential mitigations, and refine defensive search strategies over time.

***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*:** Bing publishes regular transparency reports (available on Reports Hub | Microsoft CSR or the Digital Safety Site). Microsoft's Reports Hub and Digital Safety Site are a publicly available sources where people 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 and Digital Safety Site contain additional transparency reports such as jurisdictional, community, privacy, and security transparency reports that people can easily access. As a result of the conduct of the SRA, each year Bing identifies specific areas for focused enhancements in the coming year.

In the current assessment cycle, Bing is enhancing its public transparency to provide clearer insight into the methodology used to assess systemic risks and to summarize the effectiveness of key mitigations where appropriate. This includes providing additional narrative context on how risks are evaluated, how mitigation strategies are selected and adjusted, and what indicators are used to monitor performance over time.

Finally, to ensure clarity for readers, public reporting will continue to delineate the scope of Bing Search transparency from reporting dedicated to other Microsoft services (e.g., Copilot or Bing Image Creator). Where those services are covered by separate transparency or incident reporting, Bing's public materials will reference those channels rather than conflate product scopes.

***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, **participating in several ERRS activations under the code for elections in the period** 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 "crises" under COCD.

***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 well-being. Partnering with independent third-party organizations empowers Bing product teams to take additional actions to promote more authoritative information.

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.

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 terrorists and violent extremists from exploiting digital platforms, Bing has access to the GIFCT's Incident Response processes, including ingesting hashes related to events 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 41 governments working to advance Internet freedom, and works closely with the Institute for Strategic Dialogue 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.

Election Preparation: Bing participated in multiple bilateral and roundtable engagements with EU member states upon invitation to exchange insights and has reviewed our internal defensive search playbook to election integrity for 2026. Bing has also supported a similar playbook for emerging crisis events, including levers like threat-intel monitoring and search sweep activations**.**

Bing is exploring opportunities to expand summarized responses for election topics; however, this capability should not be described as broadly deployed or available across all markets. Election-related public safety messages and answer experiences may depend on verified local content, trusted partnerships, market-specific legal and policy considerations, and the availability of authoritative sources. Accordingly, Bing should describe these mitigations as targeted and risk-based rather than universal across all elections, countries, or languages.

***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. In line with increasing regulatory emphasis on demonstrating the effectiveness of mitigations and clarity of methodology, Bing relies on independent third-party assessments such as GNI audits as part of its transparency posture and continuous improvement.

***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 people 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 people greater insight and authority over their digital environment.

Policies and Other Mechanisms to Counter Harms

***Microsoft Privacy Statement*:** People 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. 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 and reflections. Bing and Microsoft representatives have also participated in multi-stakeholder and bilateral European Elections roundtables and meetings, and have provided a post-election review report to support information-sharing and reflection on election-related reports.

In 2026, Bing continued to mature its election-related review and response processes through DIRT-led coordination, revised elections playbook review, and participation in established escalation channels.

Bing also strengthened its ability to conduct event- and election-related monitoring by developing an end-to-end capability to generate topic-based query sets within approximately one day. This capability supports faster review of emerging election-related risks and enables more timely identification of potential search quality, ranking, or defensive search issues during high-risk events.

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, policy questions, and emerging trends).

Bing's compliance function is responsible for overseeing Bing's compliance with obligations under the Digital Services Act (DSA). This includes monitoring compliance with DSA requirements, informing and advising relevant stakeholders on DSA obligations, overseeing activities related to systemic risk management and independent audits, and coordinating compliance governance and reporting. The compliance function also oversees privacy, security, and accessibility reviews and related policies, working with centralized Microsoft privacy, security, and accessibility teams responsible for company-wide policies and operations.

Bing's customer service and support function is accountable for reviewing and triaging user reports.

Bing's digital safety function is accountable for search principles of 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 labeling 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 leadership and feature teams; when regressions are observed, teams conduct issue identification and implement fixes or drive improvements.

DDR measurement is also a part of pre-launch assessment.

Bing conducts analysis on DDR for each DSA systemic risk as an underlying analysis for the formal risk assessment.

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 by 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 election queries were grounded in web results following Bing's authority-driven ranking.

Bing measures search page defect rates on periodically refreshed adversarial election-related query sets to identify potential quality, safety, and policy compliance issues.

Bing has implemented an end-to-end auto-evaluation framework for election and emerging crisis event queries, with monthly metric runs conducted for select elevated-risk elections to identify search defects and inform mitigation efforts.

Bing is also developing additional grounding metrics for summarized answers to further assess the quality and reliability of AI-generated responses.

Bing monitors jailbreak attempts in Copilot Search.

Bing also uses social listening to evaluate online conversation around potential responsible AI risks regarding election and political topics to enable mitigation actions and may carry out periodic search sweeps or red-team efforts ahead of elevated risk elections.

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 sets. 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 allow 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.
COCD	Code of Conduct on Disinformation, a set of guidelines and standards aimed at combating disinformation and ensuring transparency and accountability on 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 Abuse Material, refers to 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, the 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, an 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, a political and economic union of 27 member states located primarily in Europe.

Acronym	Definition
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, an organization, in which Bing has a leadership role, dedicated to 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, nonprofit 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, the regulatory and competition authority for the broadcasting, telecommunications, and postal industries in the United Kingdom.

Acronym	Definition
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 nonprofit 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", is a message in the public interest disseminated without charge, with the objective of raising awareness, changing public attitudes, and behavior toward a social issue.
QC	Quality and Credibility score, a metric used by Bing to evaluate and rank websites. This score helps support 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, online platforms that reach 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, online search engines that reach 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) supports efforts to mitigate risks under the DSA related to illegal content and activities and mental and physical well-being. Bing engaged with the UN INCB, a treaty-mandated body under 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 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.

The 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.

To further inform its DSA SRA methodology and fundamental-rights analysis, Bing participated in the Rights and Risk Forum held in Dublin in April 2026. The forum focused on making sure that the protection of fundamental rights, particularly freedom of expression and privacy, remains central to digital safety risk assessment methodology and practice. The discussion brought together representatives from GNI and DTSP member companies, civil society organizations, and academic experts to exchange perspectives on how technology companies can assess and mitigate risks to fundamental rights while accounting for responsible business conduct guidance, diverse regulatory expectations, trust barriers, and practical implementation considerations. Bing's participation, including engagement in a plenary discussion on company risk management, provided additional external input into its approach to risk identification, mitigation design, and evaluation across DSA risk areas, particularly those involving fundamental-rights trade-offs between user safety, privacy, and access to information.

The Lucy Faithfull Foundation, UK (LFF) supports Bing's efforts to address the Children's Harms category, including risks related to CSEAI and other forms of child exploitation online. Bing continued to work with LFF in 2026, building on prior collaboration related to Bing's CSEAI public service announcement (PSA) in the UK. This engagement included continued exchange of insights to assess and improve user interactions with the PSA and explore further enhancements, including deterrence-focused approaches. The partnership provides valuable external expertise to help Bing refine its child safety interventions, strengthen defensive measures targeting child exploitation content, and support ongoing improvements aligned with Bing's broader protections for minors.

The Tech Coalition supports Bing's efforts to address the Children's Harms category, including risks related to online child sexual exploitation and abuse. Microsoft is a founding and active member of the Tech Coalition, which facilitates industry cooperation and information exchange focused on child safety. Through Microsoft's participation, Bing benefits from cross-industry insights, researcher briefings, and working group discussions that help inform its understanding of emerging children's harms, age-appropriate digital experiences, and effective mitigation practices. This engagement complements Bing's existing child safety controls, including SafeSearch, CSEAI-related interventions, user-facing support and reporting pathways, and incident response processes. Insights from the Tech Coalition help support Bing's ongoing refinement of child safety governance, risk assessment inputs, and mitigation planning across the service.

Praesidio Safeguarding (Praesidio) supports Bing's efforts to address the Children's Harms category, including risks related to minors' exposure to harmful content, age-appropriate digital experiences, and broader online safety risks affecting children. Bing leveraged Microsoft's engagement with Praesidio in 2026, including specialist input on child safety, safeguarding assessments, and digital risk evaluations affecting minors in online environments. This engagement provided external insights into how young people experience age assurance, digital safety controls, privacy expectations, and access-to-information trade-offs across different markets. The partnership provides valuable child safety expertise to help Bing refine its understanding of emerging risks to minors, inform age-appropriate experience considerations, and support ongoing improvements aligned with Bing's broader protections for children. 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 SRA approach and supported alignment with regulatory expectations on transparency and accountability, addressing all risk areas.

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 SRA 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.

During 2026, Bing continued external engagement with Digital Services Coordinators, election authorities, and the European Commission on DSA-related risk areas, including civic discourse and electoral processes, researcher access, generative AI in search, age assurance, scams and fraud, and the treatment of illegal or harmful content. These engagements provided additional external input into Bing's assessment of evolving systemic risks and informed ongoing improvements to governance, documentation, and mitigation planning.

Bing engaged with national regulators and election authorities in connection with elections in the Netherlands, Ireland, Portugal, Hungary, Sweden, Cyprus, and Malta. These engagements included written responses, bilateral meetings, roundtables where appropriate, and requests for election-related points of contact. In April 2026, Bing responded to a request from the radio-television and Digital Services Authority in Cyprus regarding identified risks and mitigation strategies for the May 2026 Cypriot elections. Bing also responded to post-election inquiries from the Netherlands Digital Services Coordinator and offered bilateral engagement in circumstances where it did not participate in broader roundtable formats. These engagements supported Bing's ongoing monitoring of election-related risk signals and its preparedness for civic integrity risks across EU markets.

Bing also continued engagement with CnaM through regular supervisory meetings. Topics included researcher access obligations, generative AI features in search, age assurance considerations, and the categorization of Microsoft services under the DSA. During these engagements, Bing discussed the operational and legal complexity of applying age assurance in the search context, including technical, proportionality, and privacy considerations, and explained its position regarding the applicability of DSA child safety obligations to search engines. CnaM also indicated that its supervision activities would expand beyond Bing to other Microsoft services, which provides broader regulatory context for Bing's DSA governance and compliance planning.

Bing's external engagement during the period also reflected increased regulatory focus on researcher access and SRA methodology. Bing met with CnaM regarding Article 40 researcher access requirements and provided information on its researcher access compliance program. Bing also considered participation in the European Commission's May 2026 roundtable on researcher access to data, which requested operational experts from designated services. These engagements

support Bing's ongoing work to evaluate appropriate data sources, tooling, and processes for responding to researcher access obligations while accounting for data minimization, privacy, and search-specific implementation challenges.